DOCUMENT RESUME

ED 179 592 TM 009 894

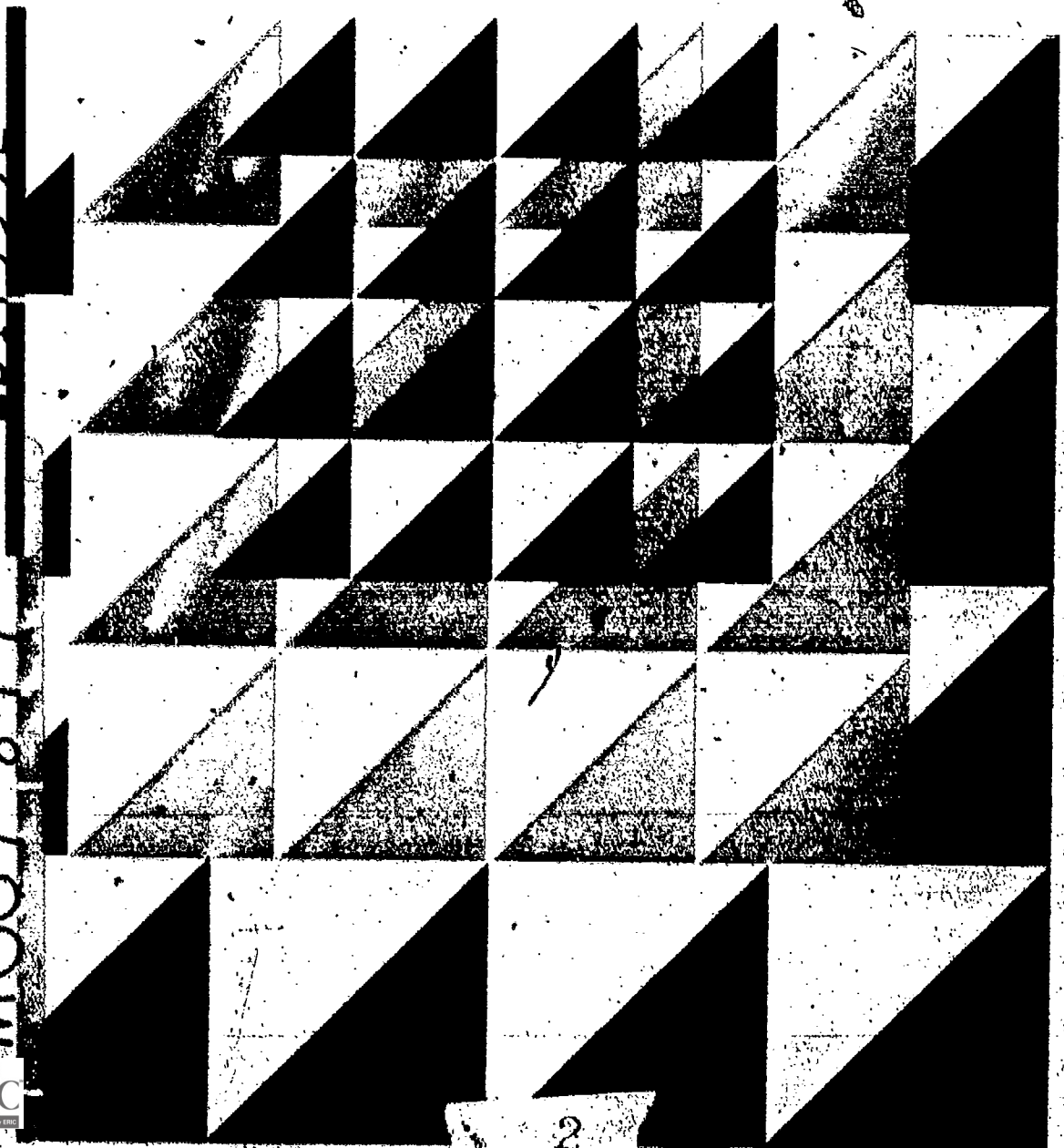
AUTHOR	Schrader, William B., Ed.
TITLE	Measurement and Educational Policy. Proceedings of the ETS Invitational Conference (Washington, D.C., October 28, 1978).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	79
NOTE	100p.
AVAILABLE FROM	Jossey-Bass, Inc., Publishers, 433 California Street, San Francisco, 94104 ($5.95)
JOURNAL CIT	New Directions for Testing and Measurement; n1 1979
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Academically Gifted; Bilingual Students; *Compensatory Education Programs; Educational Accountability; *Educational Assessment; Educationally Disadvantaged; *Educational Policy; *Educational Testing; Evaluation Needs; *Federal Legislation; Graduation Requirements; Handicapped Students; Measurement Goals; Minimum Competency Testing; Policy Formation; Program Evaluation; Resource Allocations; State Programs; *Testing Problems; Test Results
IDENTIFIERS	Connecticut; Education for All Handicapped Children Act; Elementary Secondary Act Title VII; Elementary Secondary Education Act Title I; Georgia; New Jersey

ABSTRACT

This conference was organized around four policy concerns: (1) the needs of handicapped, gifted, and bilingual students; (2) the use of test results to allocate federal compensatory education funds; (3) the validity of minimum competency testing; and (4) the demand for increasingly sophisticated evaluations. Garry L. McDaniels discussed the Education for All Handicapped Children Act; James J. Gallagher presented issues on the identification and education of gifted children; and Maria Medina Swanson described progress made in bilingual education since the Bilingual Education Act and the Lau v Nichols court decision. Joel S. Berke introduced the topic of funding allocations; Fred E. Burke presented New Jersey's rationale for combining test scores and socioeconomic status, as an index for compensatory program funding; and George F. Madaus supported the use of statewide norm-referenced achievement tests. Mark R. Shedd and R. Robert Rentz described minimum competency testing programs in Connecticut high schools and in the Georgia state colleges, respectively. Finally, Peter H. Rossi stressed the need for recognizing when not to evaluate, the difference between pilot and full-scale programs, and complications in calculating cost-effectiveness; and John Ellis illustrated how evaluations have influenced congressional appropriations. The 1978 Educational Testing Service measurement award was presented to John C. Flanagan. (CP)

directions for testing and measurement

measurement and educational policy



new directions for testing and measurement

a quarterly sourcebook
William B. Schrader, Editor-in-Chief

number 1, 1979

measurement and educational policy

proceedings of the 1978 ETS invitational conference

william b. schrader
editor

Jossey-Bass Inc., Publishers
San Francisco • Washington • London

MEASUREMENT AND EDUCATIONAL POLICY
New Directions for Testing and Measurement
Number 1, 1979
William B. Schrader, Editor

New Directions for Testing and Measurement is published quarterly by Jossey-Bass Inc., Publishers. Subscriptions are available at the regular rate for institutions, libraries, and agencies of $25 for one year. Individuals may subscribe at the special professional rate of $15 for one year. Application to mail at second-class postage rates is pending at San Francisco, California, and at additional mailing offices.

Correspondence:
Subscriptions, single-issue orders, change of address notices, undelivered copies, and other correspondence should be sent to *New Directions* Subscriptions, Jossey-Bass Inc., Publishers, 433 California Street, San Francisco, California 94104.
Editorial correspondence should be sent to the Editor-in-Chief, William B. Schrader, ETS, Princeton, New Jersey 08541.

Library of Congress Catalogue Card Number LC 78-73940

Cover design by Willi Baum
Manufactured in the United States of America

contents

introduction

Demands by educational policy makers for applications of measurement to significant new tasks are having far reaching effects on education and measurement. Measurement professionals are being asked to help in designing educational programs for, among others, handicapped children, gifted children, and bilingual children. These professionals are also being asked how measurement can help in allocating funds to schools, determining qualifications for high school diplomas and college degrees, and evaluating the worth of new educational programs. These new and complex demands have given rise to congressional debate, federal and state conferences, and extensive discussion and developmental work by measurement specialists. Measurement and educational policy is the theme of this volume, which includes the ten papers presented at the 1978 Educational Testing Service Invitational Conference.

Current educational policy is characterized by concern with the needs of special student groups. The first three chapters by Garry R. McDaniels, James J. Gallagher, and Maria Medina Swanson on this topic consider handicapped students, gifted students, and bilingual students. Each of these groups of students presents a different set of challenges to existing measurement capabilities. These chapters indicate that progress is being made in meeting these challenges and the critical next steps that need to be taken are identified.

Because funding is of central importance in the operation of schools, the possibility that test data might constitute a useful component in formulas for allocating educational funds is currently the subject of vigorous discussion. The chapters by Joel S. Berke, Fred E. Burke, and George F. Madaus on this subject point out pitfalls and safeguards in this use of tests based on both measurement and policy considerations.

The use of tests for evaluating and certifying achievement has a long and honorable history. What is new is a strong movement toward developing and using statewide minimum competency tests for high school students. At the college level, there has been a long-term trend toward greater structuring of state-supported systems of higher education. New developments in statewide testing of high school students and of college students are discussed in separate chapters by Mark R. Shedd and R. Robert Rentz.

Perhaps the most pervasive relationship between measurement

and educational policy arises from the close association between program evaluation and measurement. The demand of policymakers for increasingly sophisticated evaluations of innovations and interventions is generating new and difficult tasks for measurement. Two aspects of evaluation are discussed in this volume. The chapter by Peter H. Rossi examines the strategy of choosing appropriate programs to evaluate so that evaluation resources may be used most effectively. The chapter by John Ellis provides an insight into the way in which a variety of considerations, including program evaluations, interact in reaching decisions about federal programs.

The program for the 1978 ETS Invitational Conference was planned by: Scarvia B. Anderson (chairperson), Joan C. Baratz, Jack R. Childress, James R. Deneen, Winton H. Manning, Samuel J. Messick, Warren W. Willingham, and Jane D. Wirsig.

The papers presented at the 1978 ETS Invitational Conference provide impressive evidence that the educational community is looking to measurement for help in coping with emerging policy questions and many able measurement people are responding admirably to these demands.

William W. Turnbull

William W. Turnbull is president of Educational Testing Service.

The citation for the 1978 ETS Award for Distinguished Service to Measurement summarizes Dr. Flanagan's many contributions as scientist, scholar, and administrator.

citation: John C. Flanagan

John C. Flanagan has been well-known to several generations of graduate students for his wide-ranging technical and scientific achievements—from innovative research techniques to psychometric derivations to seminal books and articles. Over the years, those same students, maturing as researchers and professionals, have come to marvel at his ability to translate scholarly work into pioneering applications of social science—certainly the hallmark of his distinguished career.

After serving several years as associate director of the Cooperative Test Service, Dr. Flanagan organized and directed, from 1941 to 1946, the Aviation Psychology Program of the Army Air Force; this program was a giant undertaking developed to apply scientific methods of psychological measurement in the selection of pilots during World War II. A demonstrable increase in predictive validity in that selection along with a decrease in aircraft accidents justify characterizing this work as one of the dramatic success stories of applied psychology.

As founder and, for most of the past thirty years, chief executive officer of American Institutes for Research (AIR), he expanded his military research experience into a wide range of social applications. Of the hundreds of projects initiated by AIR under his leadership, perhaps the most significant was Project TALENT, the first comprehensive longitudinal study of educational development. That work led to Project PLAN, the first comprehensive computer-based program for prescribing, monitoring, and evaluating the learning progress of individual students throughout an entire school system. Finally, it is char-

acteristic of John Flanagan's vision and intellectual breadth that his attention has turned most recently to the quality of life in American society; his chief professional concerns now are how to assess the quality of life and how to develop educational and social strategies for its improvement.

It is no surprise that John Flanagan has received many honors and citations. In addition, his professional leadership has been recognized by his colleagues; he has been elected president or sectional vice president of a number of professional organizations, including the American Educational Research Association, the American Association for the Advancement of Science, the National Council on Measurement in Education, the Psychometric Society, and four different divisions of the American Psychological Association.

For his many contributions to the theory and practice of educational research and measurement, and for his productive career as scientist, scholar, and administrator, ETS has the honor to present the 1978 Award for Distinguished Service to Measurement to John Flanagan.

previous recipients of the ETS Measurement Award

1970 E. F. Lindquist

1971 Lee J. Cronbach

1972 Robert L. Thorndike

1973 Oscar L. Buros

1974 J. P. Guilford

1975 Harold Gulliksen

1976 Ralph Winfred Tyler

1977 Anne Anastasi

Legislation to meet the needs of handicapped children may bring about massive upgrading of our use of existing measurement technology.

assessing handicapped students: beyond idenfitication

garry l. mcdaniels

The Education for All Handicapped Children Act (Public Law 94-142) was signed into law by President Ford in 1975; it was to be implemented by September 1, 1978, for children between the ages of three and eighteen and by September 1, 1980, for children between the ages of three and twenty-one. This act requires that children be assessed in order to determine whether or not they are handicapped and to provide data for developing the individualized educational programs that they need. The measurement community should anticipate new demands on both its technology and its human resources as a result of the act. The purpose of this chapter is to identify two areas of weakness that may be uncovered by those demands; the weaknesses, in turn, suggest some new directions for the measurement of children over the next five to ten years.

establishing measurement guidelines

The creators of Public Law 94-142 assumed that there was a well-trained professional capacity in the United States in the area of measurement. They also assumed that this capacity was large enough and distributed widely enough to reach most of the children and youth

affected by the act. These assumptions must have existed, for in one of its sections (P.L. 94-142: Sec. 612 (2) (c)) the act required the states to institute procedures assuring that:

> All children residing in the state who are handicapped, regardless of the severity of their handicap, and who are in need of special education and related services are identified, located, and evaluated.

These children number in the millions and reside in all areas of the United States.

This basic assumption is reasonable. For example, one of the great accomplishments of psychologists during World War I (that was greatly expanded during World War II) was the creation of the large-scale testing program. The group test, the paper-and-pencil format, and machine scoring were technological breakthroughs that provided highly trained psychologists with numerous assistants and thus relieved them of direct contact with soldiers except in unusual cases. The huge screening program of the military could not have been carried out using individual assessment technology.

The measurement innovations developed in the middle of this century are commonplace in the United States today. Civilian uses of measurement devices are extensive in both schools and businesses, and the civilian work force created to administer those measurement devices is large. In addition, there is hardly a college or university in the country that does not offer numerous courses in testing and measurement. Nonetheless, the creators of the Education for All Handicapped Children Act were concerned about the abilities of the professional community. The states were directed to establish:

> Procedures to assure that testing and evaluation materials and procedures utilized for the purposes of evaluation and placement of handicapped children will be selected and administered in the child's native language or mode of communication, unless it clearly is not feasible to do so.

In addition, the lawmakers directed that "no single procedure [should] be the sole criterion for determining an appropriate educational program for a child" (P.L. 94-142: Sec. 612 (5) (C)).

The act also requires that the work of psychologists be made more public. A provision of the legislation requires that the data used for child assessment and placement be open to inspection by parents or guardians. A procedural safeguard that must be assumed by the states

(P.L. 94-142: Sec. 615 (B) (1)(A)) provides: "an opportunity for the parents of a handicapped child to examine all relevant records with respect to the identification, evaluation, and educational placement of [the] child and the provision of a free appropriate public education to such [a] child, and to obtain an independent educational evaluation of the child." This should be a boon for those measurement experts and skilled counselors who have tried for years to encourage parents to examine their children's test results and to use such results as an aid in planning academic and vocational activities for them. If the parents and guardians wish to challenge the interpretations of the available information, they are welcome to do so; the work of the measurement community is open to inspection and challenge.

No pattern has yet appeared in the problems encountered by the measurement community in implementing this act. Some anecdotal evidence, however, identifies two possible areas of weakness: the selection and administration of measurement instruments and the use of data in developing individual education plans.

Selecting and Administering Instruments. The lack of competence in the selection and administration of instruments can be illustrated by several examples. For instance, a consultant in special education who has measurement expertise told us, "I conduct workshops with school psychologists and I ask: 'How many of you use standardized tests?' All hands go up. Then I ask: 'On what groups were these tests standardized?' No hands go up." We also have reports that people are altering standardized test procedures for various disabilities with no regard for the accompanying need to modify the published norms. And there are some reports that people are using assessment devices that have low reliabilities.

It is perhaps too early to say that such isolated events constitute a pattern. There are, however, definite problems. In 1978, a panel was called together by the Bureau of Education for the Handicapped to develop criteria for evaluating the quality of local education agency assessment programs. There was consensus among this group that adequate principles now exist which, if implemented in practice, would make substantial progress in eliminating the measurement problems most frequently cited.

Given the reasonably well-developed technology of assessment and the extensive institutional training available to those pursuing careers in measurement, such patterns should not develop. A renewed commitment is needed from the measurement community to widely publicize the standards of its profession. And, if problems persist, more training for more people may also be needed, possibly supplemented by sanctions for unprofessional performance.

Developing Individual Education Programs. The lawmakers believe in the heterogeneous nature of children, and they believe that the measurement community has the capability to document their idiosyncratic characteristics. This assumption of heterogeneity underlies the mandate to develop individual education programs.

Although Public Law 94-142 asks that children receiving services be counted in one of eleven categories (with specific learning disabilities, visually impaired, deaf, and so on), these characterizations have no function beyond identifying handicapped children. In fact, some states have dropped the characterizing definitions (these include Louisiana, Massachusetts, South Dakota, and Wyoming). However, three major problems have arisen as a result of these counting categories. First, some assessment strategies may not go beyond affirming or rejecting a child's eligibility for inclusion in a certain category—a way of testing teachers' suspicions that certain children belong in particular categories. To reduce the occurrence of such situations, the Regulations* require that:

> The child is assessed in all areas related to the suspected disability, including, where appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities.

The second problem concerns the relationship between identifying handicapped children and determining their educational needs. Simply confirming a child's etiological characterization does little to help educational personnel develop a program for that child. Confirming a child's loss of hearing, for example, does little to delineate the idiosyncratic educational needs of that child. Measurement experts have discussed this paradox at great length in this decade. The criterion-referenced instrument is increasingly being recommended as a means for documenting a student's competency in specific skills, but practice in using this kind of instrument appears to be lagging behind its advocacy.

Third, and perhaps of more current concern, some standard "treatments" have become associated with various etiological characterizations of children. Some attempts to assess a child's needs seem to be dictated by initially assuming that this person is somehow homogeneous with up to a million other children in the counting category. Parents of some autistic children have reported to us, for example, that

*Rules and Regulations, Education of Handicapped Children, Implementation of Part B of the Education of the Handicapped Act. Washington, D.C.: *Federal Register*, August 23, 1977, Part II [121A.532 (F)].

since their children have been grouped for counting purposes under "emotional disturbance," the treatment strategies have been primarily psychological in character. As a result, the child's needs are assessed by a psychologist or psychiatrist who anticipates a dynamic treatment; thus they become identified with the concepts and language of dynamic psychology, which may not lead to the kind of treatment they really need.

The strength of the assumption of heterogeneity will undoubtedly be seen in court cases in the next decade. For instance, a suit is currently being brought against a local school district by the Michigan Association for Retarded Citizens. This class action alleges that the defendant has failed to provide institutionalized special education because the education provided has been directed toward chronological groups rather than toward students' individual needs. Obviously, the measurement community cannot be driven by an assumption that views children and youth who are handicapped as homogeneous. The data collected on a child's condition or needs cannot be restricted by *a priori* decisions about the category in which the child might be placed or about the treatment possibilities that exist.

conclusion

The Education for All Handicapped Children Act represents a significant challenge and opportunity for the measurement community. Rather than calling for improvements in measurement technology, the major issues in implementing the act relate to measurement practice and personnel training. However, the technical skills of measurement personnel will be on display as a result of Public Law 94-142; the measurement strategies they employ will have to respond to the assumption on which the act is based – that children are heterogeneous. That kind of response will require thoughtful, competent professionals. Thus, the new direction in measurement may be a movement toward the massive upgrading of our capacity to utilize existing technology.

Garry L. McDaniels is director, Division of Innovation and Development, Bureau of Education for the Handicapped, U.S. Office of Education.

New instruments, increased research funding, and better ways of taking account of differences in environments are necessary if we are to identify and serve the educational needs of the gifted and talented in our society.

measurement issues in programs for gifted students

james j. gallagher

The future shape of education for gifted and talented children in the United States depends on a number of factors: the ability of educators to conceptualize the special needs of and program adaptations for these children; the ability to demonstrate and evaluate meaningful progress in special programs for the gifted; developments in the rest of the educational system (desegregation, accountability, and so on); and the attitude of the general public about the desirability and importance of special education for the gifted. In this chapter, I shall examine some critical measurement issues that influence the future course of such education efforts.

What individual communities and American society as a whole decide to do about providing special educational experiences for gifted children probably depends more on societal attitudes and values than on educational innovations. Gallagher (1976) identified four broad forces that are alive in our society and that have influenced such action in the past:

1. *Egalitarianism.* There is a strong belief in the need to give all citizens equal treatment and equal opportunity and a related determination that there be no "special privileges for special people." Such attitudes, narrowly applied, can hinder special provisions for gifted

children, especially since "equal education" often gets translated into "identical education."

2. *Universal Education*. The commitment of the United States to full education for all children through high school has kept many children of limited ability in school. That situation has created a range of talents and achievement at junior and senior high school ages that is difficult to manage within a single classroom. Much of the pressure for special provisions for the gifted is a recognition of such extraordinary student diversity and the problems it creates for the conscientious teacher.

Decentralization of Educational Decision Making. When each separate school district makes its own major educational decisions, the need for special education for the gifted does not seem as pressing as other, more immediate needs. There is greater opportunity for taking a longer-range societal view at the state and federal levels. The program stimulus for the gifted often comes from those levels of government.

4. *Sense of Societal Confidence*. As long as there is overconfidence in the ability of the United States to conquer any obstacle or solve any problem as it arises, then the pressure to provide special educational help for the talented is quite low. When some of that overconfidence is lost, then there is increased pressure to build programs that would enhance the education of the most talented students in the society (and thus enhance our overall ability to meet and overcome crises).

Gifted education has profited, ironically, from World War II, the Sputnik crisis, and the current problems of energy, population pollution, and international conflict. Recognition of the social forces that influence or determine our educational policies is the first step toward understanding the otherwise curious reluctance of the society to do more for the gifted student.

gifted education in America

In the long history of Western man, we have honored many gifted individuals who provided us with new perceptions of humanity and of our environment. Plato, Mendel, Copernicus, Freud, Darwin, Curie, Shakespeare, Bronte, and Piaget have each shown us a different portrait of ourselves and our world. Those changing portraits, in turn, have resulted in major transformations both in our society and in civilization. And although we often do not recognize them, below this level of genius are layers of other gifted and talented individuals who have made significant, although less society-shaking, contributions. The scientific discoveries, the creative writing, the art, and the music that this second echelon of gifted individuals has produced have also played a major role in changing the total fabric of our civilization.

Ignoring the education of these gifted and talented individuals cheats both them and the larger society of their true potential. Yet we hesitate when considering special education provisions for gifted and talented children, and we listen to counsel suggesting that the gifted will make their contributions without any special educational aid or help. A strong case can be made for the presence of a love-hate relationship between giftedness or talent and American society. On the one hand, we revere the gifted individual who has risen from a humble background. We are proud to live in a society where talent can triumph over environment or family status. But on the other hand, since our origins came from battling an aristocratic elite, we are suspicious of attempts to subvert our commitment to egalitarianism. We do not want a new elite to develop; as a result, we waver in our attitudes. We design our elementary and secondary programs for gifted students in ways that can be defended by cautious administrators as giving no special favors and not tipping the scales in favor of the societally powerful or specially endowed (Gardner, 1961).

Kurt Vonnegut, Jr. (1950, p. 7) has carried one of the common feelings about the gifted in our society to a logical conclusion in his short story, *Harrison Bergeron*, which is set in some future society:

> The year was 2081, and everybody was finally equal. They weren't only equal before God and the law, they were equal in every which way. Nobody was smarter than anybody. No one was better looking than anybody else.

The reason for this enforced equality was that people who were outstanding in various ways were given handicaps to equalize the society. There was a government agency, headed by the Handicapper General, whose job it was to enforce such equality. Those citizens who could dance well had to wear sandbags on their feet; those who were strikingly good-looking had to wear masks so as not to embarrass those who were not. And what about those with high intellectual ability?

> George, while his intelligence was way above normal, had a little mental handicap radio in his ear. He was required by law to wear it at all times. Every twenty seconds or so, the transmitter would send out some noise to keep people like George from taking unfair advantage of their brains.

The essentially destructive approach to "equality" satirized by Vonnegut influences our feelings about the gifted until we reach higher education, when a miraculous transformation takes place.

The United States has created the most complex and extensive

higher education and professional school establishment in the world. We may not think of the curricular offerings of the Stanford Medical School or the Harvard Law School as programs for gifted students, but we know that they are; and no apologies are made for the fact that only the "best" students are allowed to attend. After all, some of us may need a good lawyer from time to time, others may need an excellent surgeon, and others would like some good advice from a competent psychiatrist.

current status

The history of support for programs for exceptional children in the U.S. Office of Education gives us some insight into the cultural problems of the gifted and the talented in our society. The federal government will provide over $900 million in fiscal year 1979 to improve the education of school-aged handicapped children. These dollars are certainly needed; in fact, they do not provide all that handicapped children need in the way of special education services. But during that same year, the federal government will provide only slightly over $3 million for gifted and talented children. In short, for every dollar spent on a gifted child for special education, $100 is spent on a handicapped child.

Is this the appropriate rate of expenditure for exceptional children in our society? Probably not. It is representative, however, of the political realities that attend our present system of crisis decision making in government. Gifted children suffer because they are a "cool," or long-range, problem. Budget and legislative decisions are made not on the basis of what might be of ultimate benefit to society but on what is the greatest immediate crisis or what represents the largest political pressure. Gifted children may be our best long-range investment in education, but they do not create problems of immediate significance; nor have they had a vocal constituency capable of extracting attention and dollars from public policy makers.

Mitchell and Erickson (1978, p. 13) report from a national survey on current policies, resources, and services that the national picture of educational programs for gifted and talented children in 1976-77 is slightly better than it was in 1971-72: More gifted and talented students are being identified and served; more states have statutes and policy documents concerning their education; more money is being allocated to educational programs for these special children; more personnel are being assigned to work in this area; and more training is available. They concluded, however, that, "Despite the fact [that] there is 'more of everything' now than there was in 1972, . . . the

United States still falls far short of meeting the educational needs of this special segment of its population." They also concluded that federal entrance into the issue of educating the gifted and talented did have one important effect; though modest in its fiscal efforts, the government modified and extended the generally accepted definition of the gifted child.

who are the gifted?

Each culture tends to define giftedness in its own image; the definition not only fixes the role of the gifted individual in a certain culture, but it tells us something about the culture itself as well. What would be called *gifted* in a primitive society may be very different from what we would honor in our advanced technological society. Some cultures, such as that of ancient Greece, honored the orator, while Rome valued the engineer and soldier, and so on. What does the current definition tell us about our own culture? According to Marland (1978, p. 10):

> Gifted and talented children are those identified by professionally qualified persons who by virtue of outstanding abilities are capable of high performance. These are children who require differentiated educational programs and services beyond those normally provided by the regular program in order to realize their contribution to self and society.
>
> Children capable of high performance include those with demonstrated achievement and/or potential ability in any of the following areas:
>
> 1. General intellectual ability
> 2. Specific academic aptitude
> 3. Creative or productive thinking
> 4. Leadership ability
> 5. Visual and performing arts
> 6. Psychomotor ability

Such a definition is a noble attempt to broaden the idea of giftedness beyond verbal facility, but it cannot become operational without adequate measuring instruments and more sophisticated theory.

measurement influences

After six decades of trying to measure individuals' characteristics, we are now engaged in an attempt to understand and predict those

individuals' future behaviors and performances. This attempt has worked reasonably well in the areas of achievement and cognitive development, but it has worked less well in such areas as creativity and leadership. Predictions of creativity and leadership depend, in large measure, on the nature of the specific environment in which an individual is behaving, as well as on the characteristics of that individual. Thus, individual X may be a potential leader in environments 7, 13, and 22 but not in 5, 8, or 9. This interactive approach to measuring leadership and creativity lacks the decisive ring of saying that someone is a "born leader," but it is probably more accurate in the long run (Arnold, 1977; Stogdill, 1974).

Cultural Differences. A special problem is encountered in identifying gifted minority group children who have grown up under different cultural circumstances than have those children assessed by standard IQ measurements. There have been three general approaches to this problem to date. The first of these can be called a statistical adjustment. Mercer (1978) has developed a technique known as the System of Multicultural Pluralistic Assessments (SOMPA). This system makes statistical adjustments for students' actual IQ scores based on the presence or absence of optimum assessment conditions. According to Mercer, optimum conditions are present if all students: (1) have had similar opportunities for learning the materials and acquiring the skills covered in the test; (2) have been similarly motivated by the significant other persons in their lives to learn this material and to acquire these skills; (3) have had similar experience with taking tests; (4) have no emotional disturbances or anxieties interfering with test performance; and (5) have no sensory-motor disabilities interfering with prior learning or with their ability to respond in this test situation. Mercer believes that when these factors are held constant the pluralistic model assumes that the individual who has learned the most probably has the greatest learning potential. Use of this technique has been successful in identifying gifted and talented minority group children who otherwise might not have been located.

A second major approach to identifying gifted minority children is to try to assess with measuring instruments the characteristics in those domains that the cultural subgroup puts particular stress on. In this way, one can identify the special talents in different ethnic groups. For instance, Bernal (1974) suggests such a test for young Chicano children based on Piagetian concepts and including the Cartoon Conservation Scale developed by DeAvila and Havassy (1975). In another example, Meeker (1978) reports the work of Evelyn Hahn of the Bureau of Indian Affairs in identifying gifted Navajo students. By using Structure of Intellect tests that are heavily weighted to *figural* rather than to

semantic areas, Hahn was able to find gifted Indian children. The Navajo children tested had particularly high scores for auditory memory, but they scored low on classification skills in the figural dimension. Navajo is a sparse language with a minimum of words for classification, and it is learned largely through the auditory sense. Thus, this type of identification provides a basis for understanding cultural differences as well as for plotting some clear curriculum objectives for Navajo children.

Torrance (1976) reports two types of special tests designed to identify giftedness in black populations. These tests are "sounds and images," and "thinking creatively with action and movement." In the "sounds and images" test, children are asked to describe images suggested by a series of sound effects. Test results indicate that black and white children have equally rich imagery storehouses. However, in the second kind of test, "thinking creatively with action and movement," Torrance found that black children responded to problems with action and movement while white children tended to respond verbally, telling rather than acting out what they would do. This test allowed Torrance and his coworkers to use the specially developed talents of the black subgroup to help identify its gifted and talented members.

The third major technique that has been used to identify gifted minority children combines tests, rating scales, and peer and adult nominations. This approach is presented in a systematic form by Baldwin (1978). She uses eleven different assessment instruments, ranging from standard intelligence tests to peer nominations, to develop a composite score for an individual. The use of multiple measures enables her to find the gifted and talented students within minority groups without unduly penalizing students for poor performance on any one of the instruments.

The identification of gifted and talented students within minority groups has progressed much more rapidly than has the development of clear and distinctive curriculum adjustments for them. Although some suggestions have been made (Gallagher and Kinney, 1974), the field still lacks definitive statements regarding important distinctive curriculum adaptations for these youngsters (Baldwin, 1978).

Creativity. Great interest in creativity was spawned by the theoretical work of Guilford (1950, 1967) and spurred by the imaginative application of that work by Getzels and Jackson (1962) and by Torrance (1965). This movement created a blizzard of new measuring instruments of dubious validity and reliability. Such simple instruments, of course, did not measure creativity, which is a complex process that cannot be viewed apart from the subject and the environment. However, they did measure some characteristics of intellectual fluency and

flexibility, which may be more matters of cognitive style than separate intellectual operations. They miss the essence of the complex process of creativity as noted in the study of the creative person (Barron, 1969).

school adaptations

The two major objectives of special education for gifted students have generally been agreed upon (see Gallagher, 1975): (1) they should master the structure of the knowledge disciplines and understand the basic principles at the heart of their subject matter. They should learn systems of knowledge rather than simple facts and associations. (2) They should learn the heuristic skills of problem solving, creativity, scientific method, and so on, so that they will become more autonomous learners and not be constrained by the limits of individual teachers. A number of adaptations have accompanied efforts to meet these long-range goals.

Content. During the early 1960s, a brief but exciting marriage between scholars and educators attempted to produce a systematic reorganization of knowledge in mathematics, physical science, and the social sciences (Bruner, 1960; Goodlad, 1964). Programs that were developed during this period emphasized the basic structure of a discipline, stressed the importance of having the student behave as a physicist, a historian, or whatever, and encouraged the introduction of complex ideas as early as possible in the school program. These are all educational goals that fit the needs of gifted children very well. This marriage disintegrated in the late 1960s when the Vietnam war and desegregation took over as major emphases for schools and scholars. However, it pointed the way toward a new liaison that can aid the clear presentation of important ideas to gifted and talented students.

Examples of how such synthesis of important ideas can be accomplished, as well as verification of the viability of the approach, have been presented by two television series produced by the BBC: Kenneth Clark's *Civilisation* (1970) and Bronowski's *Ascent of Man* (1973). Each series tried to take central ideas and major insights and build a set of illustrative examples, conceptual linkages, and consequences around them. A few brief quotes from the Bronowski series will illustrate major ideas that are well within the grasp of the gifted and talented from preadolescence onward.

> War, organized war, is not a human instinct. It is a highly planned and cooperative theft. And that form of theft began ten thousand years ago when the harvesters of wheat accumulated a surplus and the noamds rose out of the desert to rob them of what they themselves could not provide (p. 88).

> The different cultures have used fire for the same purposes: to keep warm, to drive off predators and clear woodland, and to make simple transformations of everyday life, to cook, to dry and harden wood, to heat and split stones. But, of course, the great transformation that helped us make our civilisation goes deeper: it is the use of fire to disclose a wholly new class of materials, the metals (p. 124).

> Easter Island is over a thousand miles from the nearest inhabited island. . . . Distances like that cannot be navigated unless you have a model of the heavens and of star positions by which to find your way. People often ask about Easter Island, how did men come here? They came here by accident: that is not the question. The question is why could they not get off? And they could not get off because they did not have a sense of the movement of the stars by which to find their way (p. 192).

> The horse and the rider have many anatomical features in common. But it is the human creature who rides the horse, and not the other way about. There is no wiring inside the brain that makes us horse riders. Riding a horse is a comparatively recent invention—less than five thousand years old. And yet it has had an immense influence, for instance, on our social structure. Plasticity of human behavior makes that possible. That is what characterizes us in our social institutions, of course, and above all, in our books, because they are the permanent products of the total interest of the human mind (p. 412).

Such ideas can be the base of an exciting curriculum if scholars and teachers renew their joint efforts and interests.

Skills. The earlier noted adventures in search of creativity and the creative process have focused attention on the thinking process and generated some useful instructional programs and materials (Feldhusen and Treffinger, 1977; Torrance and Myers, 1970).

Learning Environment. Several innovative administrative devices have been adopted in education for gifted students, such as special schools, magnet schools, resource rooms, mentors, and tutorial programs; they are all designed to create an environment conducive to achieving the two major objectives of such special education. But evaluating educational programs for the gifted has been difficult without appropriate measuring instruments, since standard achievement tests leave much to be desired in this regard (Renzulli, 1976). Since multiple-choice achievement tests must be constructed such that they allow most of the students to respond to each item, there is no room on the test for

the kind of knowledge that *only* the gifted child might learn or understand.

Consequently, despite the high scores that gifted students obtain on standard norm-referenced achievement tests, one still may overlook their real capabilities. For example, standard achievement tests in history stress much factual knowledge and some reasoning ability. Such tests may indicate whether or not a youngster has necessary information regarding the American Revolution or the U.S. Constitution, but they are unlikely to demonstrate the gifted child's understanding of revolution as a generic concept and his or her ability to apply that knowledge to a wide variety of circumstances. Educators would not be able to discern from multiple-choice testing of simple concepts in astronomy or physics or from science achievement tests that a youngster has a grasp of Einstein's theory of relativity. It would be counterproductive to place sophisticated items like this on a standard multiple-choice achievement test; the vast majority of students would miss them totally, which would create problems in test construction and in norming.

conclusion

If there were no interest in doing something special or unique for gifted students, there would be no need to think about better or different measuring instruments. But if we were correct in our original assumption that crisis heightens our appreciation of gifted students and their needs, then we probably can be confident, looking at our immediate future, that the need for special programs (and thus for special instruments) will be recognized.

We must discard standard instruments designed for average students and develop instruments for special populations and unique educational objectives; a very special and unique type of criterion-referenced test is needed – one that is designed to measure maximum rather than minimum competence. In addition, a new set of instruments would allow us to integrate knowledge of the individual with the classification of common environmental settings and conditions. This would help us to properly identify students for leadership and creativity programs and to find hidden talent in minority groups (Baldwin, Gear, and Lucito, 1978).

The chronic absence of research and training money for evaluating and teaching gifted children has led to a disastrous lack of interest on the part of universities in this topic. The crass financial truth is that training programs with few enrollees, such as education for the gifted, cannot pay for themselves and must have external support if universities are to become involved. And since most innovative ideas in

education still come either from universities or from the research community, some degree of incentive must be provided if we are to see dramatic innovation in programs for gifted children in the near future. Moreover, there is a lack of any deliberate policy to encourage the development of new instruments for special purposes. Instrument development is not considered an appropriate use of limited research dollars. Unfortunately, *research* cannot compete politically with *service* for scarce funds. Service programs provide direct benefits to their constituents and create instant political rewards, but research may not have measureable impact until several generations of politicians have passed by. One solution that has become more and more popular among scientists concerned with public support has been proposing that a fixed percentage of service funds be set aside for research and development. In this way, research and development activities would become political beneficiaries of pressures for increased service. As Gallagher (1975, p. 26) says;

> One alternative to current operations would link priority programs to some sliding scale related to general education expenditures. For example, educational research and development could be tied to educational expenditures and receive five percent of the total, whatever that total is. The more money spent on educational services, the more money would go to research, and at a percentage level shown to be effective in fields such as agriculture and health. This would eliminate the temptation for budget cutters looking for lost dollars to attack a program whose nature makes it more defenseless than programs with strong emotional support, such as programs for services to the handicapped.

Other observers of the federal scene have proposed similar schemes. For example, Challoner (1974) has suggested the formation of a biomedical research trust fund that would be tied to the gross revenues of the health industry or perhaps to a percentage of health insurance premiums. And Krathwohl (1977) has suggested that a fixed percentage of the federal education allocation go to educational research and development. Unless some such system-wide strategy is adopted that will allow long-range goals of great merit, such as the education of gifted students or the discovery and development of new ideas in measurement to be supported or underwritten, we must continue to live with the less than optimum level of support that now exists.

We in education seem to support the philosophy that new measuring instruments appear as if by magic – perhaps through a firm tap

by a fairy godmother. Maybe federal agencies are afraid they will be attacked for trying to subtly influence the curriculum from a national standpoint. Whatever the reason, education for the gifted and talented has suffered substantially from having to put on a suit of measurement clothes that neither fits its needs nor measures its intellectual breadth. Obviously, the pitiful sum designated in the federal budget for educating the gifted and talented is not sufficient for research, training, or instrument development. We may have to rely on private sources, such as foundations, for the statesmanship and the foresight needed to support new measurement and program innovations. Science flies on the wings of its measuring instruments; until we, as a nation, recognize and act on that fundamental concept, our vision of what is possible for the gifted student will be limited by our own inadequate instruments. The prevailing viewpoint of those who support special programs for the gifted is summed up in a quote from Arnold Toynbee (1968, p. 24):

> The creator has withheld from man the shark's teeth, the bird's wings, the elephant's trunk, and the hound's or horse's racing feet. The creative power planted in a minority of mankind has to do duty for all the marvelous physical assets that are built into every specimen of man's nonhuman fellow creatures. If society fails to make the most of this one human asset, or if, worse still, it perversely sets itself to stifle it, man is throwing away his birthright of being the lord of creation and is condemning himself to be, instead, the least effective species on the face of this planet.

references

Arnold, A. "Leadership — A Survey of Literature." In A. Arnold (Ed.). *A New Generation of Leadership*. Ventura, Calif.: National/State Leadership Training Institute of the Gifted and Talented, 1977.

Baldwin, A. "The Baldwin Identification Matrix." In A. Baldwin, G. Gear, and L. Lucito (Eds.). *Educational Planning for the Gifted: Overcoming Cultural, Geographic, and Socioeconomic Barriers*. Reston, Va.: Council for Exceptional Children, 1978.

Baldwin, A., Gear, G., and Lucito, L. (Eds.). *Educational Planning for the Gifted: Overcoming Cultural, Geographic, and Socioeconomic Barriers*. Reston, Va.: Council for Exceptional Children, 1978.

Barron, F. *Creative Person and Creative Process*. New York: Holt, Rinehart and Winston, 1969.

Bernal, E. "Gifted Mexican American Children in Ethno-Scientific Perspective." *California Journal of Educational Research*, 1974, *25*, 261–273.

Bronowski, J. *Ascent of Man*. Boston: Little, Brown, 1973.

Bruner, J. *The Process of Education*. Cambridge, Mass.: Harvard University Press, 1960.

Challoner, D. "A Policy for Investment in Biomedical Research." *Science*, 1974, *186*, 27–30.

Clark, K. *Civilisation*. New York: Ronald Press, 1970.
DeAvila, E., and Havassy, B. "Piagetian Alternatives to IQ: Mexican-American Study." In N. Hobbs (Ed.), *Issues in the Classification of Children: A Sourcebook on Categories, Labels, and Their Consequences*. San Francisco: Jossey-Bass, 1975.
Feldhusen, J., and Treffinger, D. *Teaching Creative Thinking and Problem Solving*. Dubuque, Iowa: Kendall/Hunt, 1977.
Gallagher, J. *Teaching the Gifted Child*. Boston: Allyn & Bacon, 1975.
Gallagher, J. "Needed: A New Partnership for the Gifted." In J. Gibson and P. Chennels (Eds.), *Gifted Children: Looking to Their Future*. London, England: Anchor Press, 1976.
Gallagher, J., and Kinney, L. (Eds.). *Talent Delayed—Talent Denied: A Conference Report*. Reston, Va.: Council for Exceptional Children, 1974.
Gardner, J. W. *Excellence: Can We Be Equal and Excellent Too?* New York: Harper & Row, 1961.
Getzels, J. W., and Jackson, P. W. *Creativity and Intelligence*. New York: Wiley, 1962.
Goodlad, J. *School Curriculum Reform in the United States*. New York: Fund for the Advancement of Education, 1964.
Guilford, J. P. "Creativity." *American Psychologist*, 1950, *5*, 444–454.
Guilford, J. *The Nature of Human Intelligence*. New York: McGraw-Hill, 1967.
Krathwohl, P. "Improving Educational Research and Development." *Educational Researcher*, 1977, *6* (4), 8–14.
Marland, S. *Education of the Gifted and Talented*. Report to the Subcommittee on Education, Committee on Labor and Public Welfare, U.S. Senate, Washington, D.C.: 1972.
Meeker, M. "Nondiscriminatory Testing Procedures to Assess Giftedness in Black, Chicano, Navajo, and Anglos." In A. Baldwin, G. Gear, and L. Lucito (Eds.), *Educational Planning for the Gifted: Overcoming Cultural, Geographic, and Socioeconomic Barriers*. Reston, Va.: Council for Exceptional Children, 1978.
Mercer, J. *SOMPA Technical Manual*. New York: The Psychological Corporation, 1978.
Mitchell, P., and Erickson, D. "The Education of Gifted and Talented Children: A Status Report." *Exceptional Children*, 1978, *45*, 12–17.
Renzulli, J. *A Guidebook for Evaluating Programs for the Gifted and Talented*. Ventura, Calif.: National/State Leadership Training Institute of the Gifted and Talented, 1976.
Stogdill, R. *Handbook of Leadership: A Survey of Theory and Research*. New York: Free Press, 1974.
Torrance, E. P. *Rewarding Creative Behavior*. Englewood Cliffs, N.J.: Prentice-Hall, 1965.
Torrance, E. P. *Thinking Creatively in Action and Movement*. Athens: Georgia Studies of Creative Behavior, University of Georgia, 1976.
Torrance, E. P., and Myers, R. *Creative Learning and Teaching*. New York: Dodd, Mead, 1970.
Toynbee, A. "Is America Neglecting her Creative Talents?" In C. Taylor (Ed.), *Creativity Across Education*. Ogden: University of Utah Press, 1968.
Vonnegut, K., Jr. *Welcome to the Monkey House*. New York: Dell, 1950.

James J. Gallagher is Kenan Professor of Education and director, Frank Porter Graham Child Development Center, University of North Carolina at Chapel Hill.

Increased understanding of the complexities of bilingual education is yielding better tests and more effective use of tests results.

testing and bilingual education

maria medina swanson

Millions of students in the United State come from homes in which a language other than English is spoken. A growing awareness of their special problems has led to the enactment of numerous federal and state laws affecting the education of such students. The impact of these changing educational policies on instructional programs, as well as on educational and psychological measurement, is being felt across the country. Yet the need to properly identify and diagnose the specific linguistic and educational needs of these non-English-speaking students in order to provide meaningful educational experiences for them while they are learning English remains a crucial issue for those involved in bilingual education.

Throughout the history of the United States, there have always been students for whom English is a second language. And throughout that time, except for a period beginning in the late 1890s and ending in the mid-1960s, many of these students have been able to enroll in schools that offer instruction both in their native language and in English (Leibowitz, 1978). "An estimated one million children attended bilingual programs in public schools during the nineteenth century, not to mention the continuing tradition which started even earlier in sectarian schools" (Zirkel, 1978, p. 48). Such programs were

available, for example, in Spanish-English public schools in New Mexico, French-English schools in Louisiana, and German-English schools in several midwestern and northeastern states.

Toward the latter part of the nineteenth century, however, due to a combination of circumstances (none of which had anything to do with educational needs)—increasing immigration, religious and ethnic prejudice, and nationalism—a wave of laws prohibiting instruction in any language other than English in public and even private schools spread from state to state. This attitude was compounded by our involvement in World War I; during those years, we pushed the xenophobic panic button. It was absolutely *verboten* to speak German, and speaking any other language was considered suspiciously un-American. Some states went so far as to levy a fine against anyone overheard speaking German in a public place. Other states tried to ban foreign language instruction altogether. The effects of this hypernationalism were far-reaching: by 1923, thirty-four states had statutes requiring English to be the only medium of instruction in public schools. Its impact lasted well into the sixties, though we still can see some vestiges of it today.

In the sixties, our country finally began to awaken. The Civil Rights Act of 1964 made us more aware than ever before of racial and ethnic minority groups and their needs, the many deprivations and injustices they suffered, and their emerging political strength. The 1960 census revealed a phenomenal growth among the Mexican-American population in the Southwest, which by then accounted for 12 percent of the total combined population of Texas, California, Colorado, Arizona, and New Mexico. In New York and other northeastern states, the influx of Puerto Ricans and other Hispanic immigrants was also cause for concern. Federal and state governments began to respond to this growing constituency: the Equal Employment Opportunity Commission studied employment patterns, the Civil Rights Commission examined legal rights, Congress suspended English literacy requirements for voting, and a number of states looked into educational issues affecting the different minority groups.

Linguistic minorities began to speak out as well. They were understandably dissatisfied with the failure of the educational system to meet the needs of their children. In far too many instances, schools would automatically place students with limited English proficiency in classes two or three grades below their age group, hoping to make it easier for them to catch on to English. The results were usually more damaging than beneficial. In other cases, such students were placed with low-ability groups at the elementary level and/or channeled into vocational programs in junior and senior high schools. In addition,

unwilling to accept the idea that in order to succeed they must give up their cultural and linguistic traditions, ethnic communities throughout the U.S. began demanding the kind of instruction that was responsive to their needs—bilingual education. Their rationale was simple and straightforward: build upon children's strengths by teaching them in their own languages while they learn English. And they had the special incentive of knowing that such programs were indeed feasible: the successful bilingual program implemented in 1963 at the Coral Way School for Cuban refugees in Dade County, Florida, had served as a model for a few innovative schools in the Southwest and had helped popularize the concept of bilingual education among ethnic communities.

The educational community also became involved in the quest. For example, in 1966, the National Education Association sponsored a conference on the education of Spanish-speaking children and in its report strongly recommended bilingual instruction. Other groups reached similar conclusions and recommended involving the federal government. Thus, the road was paved for the Bilingual Education Act.

Title VII: the bilingual education act

In 1968, Congress took positive steps to help children who could not understand instruction in English. Title VII (the Bilingual Education Act) of the Elementary and Secondary Education Act (cited in Schneider, 1976, p. 172) included this declaration of policy:

> In recognition of the special educational needs of the large numbers of children of limited English-speaking abilty in the United States, Congress hereby declares it to be the policy of the United States to provide financial assistance to local educational agencies to develop and carry out new and imaginative elementary and secondary school programs designed to meet these special educational needs. For the purposes of this Title, "children of limited English-speaking ability" means children who come from environments where the dominant language is other than English.

At last, the "sink or swim" approach, which had contributed to a high dropout rate among Hispanics and students from other linguistic minorities, was recognized as ineffective and the English only policy was overruled. This was truly landmark legislation. In order to provide "new and imaginative" programs, it authorized such activities as: (1) bilingual education programs; (2) programs designed to teach students about the history and culture associated with their languages;

(3) efforts to establish closer cooperation between school and home (4) early childhood education programs; (5) adult education programs (for parents of students); (6) programs for dropouts or potent dropouts in need of bilingual instruction; and (7) programs conducted by accredited trade, vocational, or technical schools.

Shortcomings. Title VII also authorized planning grants, research grants, and pilot projects to test the plans as well as the development and dissemination of the bilingual instructional material. And funds were made available for preservice and inservice training of a variety of instructional and ancillary personnel (Schneider, 1976). The act, however, had a few shortcomings. Most noticeable among them was the absence of a definition of bilingual education. This was remedied in the manual published by the Office of Education (U.S. Office of Education, 1971):

> Bilingual Education is the use of two languages, one of which is English, as mediums of instruction for the same pupil population in a well-organized program which encompasses part or all of the curriculum and includes the study of the history and culture associated with the mother tongue. A complete program develops and maintains the children's self-esteem and a legitimate pride in both cultures.

Another shortcoming was the "poverty clause" requiring that participating students be from families that earned less than $3,000 annually or were on welfare. This limitation was removed in amendments made in 1972. But what was actually the greatest drawback of all was the general lack of experience of all personnel involved in implementing the Bilingual Education Act and the scarcity of outside experts to provide the necessary technical assistance. In order to implement the kinds of programs called for in the act's guidelines, personnel would have to be able to conduct linguistic and educational needs assessments, population studies, and community surveys; design and plan programs, including long-range goals and five-year program objectives; design instructional components with process and product objectives in first and second languages, content areas, and culture and heritage (including procedures for evaluation, data collection, analysis, and reporting); acquire, adapt, and develop instructional materials for student use as well as training materials for staff development; design and conduct a staff development program for teachers, paraprofessionals, and support personnel; conduct a program evaluation outlining behaviors to be measured, instruments to be used, methods of data collection, and methods of analysis; and involve

parents and community in school activities and advisory councils and design adult programs for them. In addition, these personnel would be responsible for teaching students, grading papers, and supervising the lunchroom.

It is not surprising, then, that in 1970 when the Office of Education commissioned the Rand Corporation to conduct a study of several of its programs, the findings showed that bilingual programs were the hardest to implement. "Title VII began with the fewest available resources and the least developed program strategy" of any of the programs, the study added (Andersson and Boyer, 1978, p. 40). The implementation problems were attributed to inadequate materials, unrealistic goals, impossible schedules, and an overburdened staff. Although the study's authors acknowledged that the relative newness of bilingual education may have been primarily to blame (the study was made only one year after the start of the bilingual program), they also observed that the changes attempted by some projects may have been too ambitious. Lack of experience may have accounted for a slow and rather painful beginning; nevertheless, the dedication and enthusiasm of the professionals committed to the philosophy of bilingual education resulted in continuous efforts to improve all aspects of the program.

By 1973, third- and fourth-year bilingual education programs showed substantial progress in program design and instruction; selection and development of materials, teacher training, and community national projects had been established to provide services for bilingual instructional programs. For example, the Materials Acquisition Project identified and evaluated published materials for bilingual instruction, and the Dissemination Center for Bilingual Bicultural Education (DCBBE) published and distributed selected project-developed materials; in this way, some of the initial demands of individual projects for development of materials were met. Progress had also been made in identifying achievement, language dominance, and language proficiency tests that could be used in bilingual programs. Many of these tests had been developed specifically for bilingual students. An annotated bibliography listing seventy-nine project-developed instruments available from noncommercial sources was published by the DCBBE (Dissemination Center for Bilingual Bicultural Education, 1975).

Assessing Students' Eligibility. The years between 1968 and 1974 made up an important learning period for bilingual educators. The method of identifying students eligible to participate in bilingual programs went through a series of developmental stages. At first it was not uncommon to find students being diagnosed as limited English-speaking and thus needing bilingual education simply on the basis of

their surnames. In other cases, such placement was based on nothing more than teachers' opinions about students' language dominance as indicated by their classroom performance in English. Dissatisfaction with these assessment procedures led to the use of a combined approach consisting of (1) a questionnaire designed to determine which language students used at home, with their peers, on the playground, and so on; (2) a language dominance test (an oral interview) during which students were asked to answer questions or to tell stories about pictures or objects in both their native language and English (or they might be asked specific questions about their homes, families, and schools, or otherwise engaged in conversation in both languages); (3) input from teachers; and (4) direct observation of students by the evaluators.

Although the combined approach generally resulted in adequate determinations of language dominance, educators eventually realized that language dominance and language proficiency were two different things and that, although language dominance determined a student's need for bilingual instruction, it told very little about the degree of that student's proficiency in either language. For instance, a third grade student transferring to the school after completing the first and second grades in Puerto Rico is obviously much more proficient in Spanish than is a third grade student whose Puerto Rican parents speak Spanish at home but who has struggled through the first and second grades using English in the United States. Both students are Spanish-dominant and both have limited English language proficiency; however, the first has a relatively rich and extensive vocabulary and can read and write in Spanish, whereas the second, although well-versed in conversational Spanish centering on family and neighborhood topics, has had far less linguistic experience than has the first. Thus, teachers soon learned that a class of thirty-five Spanish-dominant students could very well mean a class with anywhere from one to thirty-five different levels of proficiency in Spanish and just as many different levels of proficiency in English, resulting in Excedrin Headache Number 70 for the teacher. Curriculum planning, materials selection and adaptation, and instructional approaches and techniques had to take these individual differences into account. Qualified teachers had to be able to not only teach content areas in two languages but be masters in individualization, small-group instruction, materials adaptation, diagnostic procedures; and above all else, they must be warm, sensitive, perceptive, and flexible.

Standardized Testing. The complexities that diverse levels of language proficiency brought to the classroom were compounded in the area of standardized testing. The need to develop instruments in the language of the students proved to be a very complicated under-

taking. Translating existing English language tests proved unsatisfactory because people of different ethnic and linguistic backgrounds do not think in the same way, structure thoughts in the same manner, or learn equivalent words and concepts in the same order. A word or concept that is common and therefore considered easy among English-speaking children may not be at all common or even exist in the same form in another language. One example of this problem is the English word *pet* (DeAvila and Havassy, 1978). There is no such word in Spanish. The usual translation is *animal domestico* or *mascota*, depending on the meaning. Both of these concepts are considerably more complicated than the English *pet*.

Obviously, special tests had to be developed for these students. Psychologists, consultants, evaluators, teachers, project directors, counselors—anyone who had a good idea—began developing tests during this period. Even one or two commercial publishers decided to give it a try. Additional problems soon surfaced: regional differences, both linguistic and cultural; lack of reading skills in the native language; and gaps of proficiency in the native language that, in many cases, were filled in the second language (English). Thus, testing a fourth grader's achievement in science, math, or social studies, for example, may have required giving instructions, questions, and answers in both the student's native language and English. And a psychological evaluation had to consider the possibility that a child might know some things only in one language and others only in the other language.

state involvement

When the Bilingual Education Act (Title VII) was enacted in 1968, twenty states still prohibited instruction in a language other than English. However, its passage brought about a surge of activity in state legislatures across the country. They passed laws to lift restrictions against the use of other languages, laws to allow bilingual instruction, and laws that appropriated moneys for bilingual programs. A number of states adopted laws requiring psychological evaluations in the child's native language and prohibiting any placement of children in special education classes until such assessment had been made. In 1972, Massachusetts became the first state to require bilingual education programs in all schools with twenty or more students of limited English-speaking ability. Soon Texas, California, Colorado, New Mexico, and Illinois followed with similar mandates. By 1976, ten states had statutes making bilingual education mandatory; sixteen states specifically permitted it; fourteen states had no statutes but tacitly allowed it; and ten states

still prohibited it in some form or other, although some of them managed to have Title VII programs in spite of such regulations (Development Associates, 1977). Among the factors that accounted for state involvement in bilingual education were the number of linguistic minority students residing in the state, the degree of political activity of the ethnic community, exposure to bilingual education through Title VII programs, and the level of awareness in the state about the need for and the implementation of bilingual education.

State involvement really intensified the flurry of activities surrounding bilingual education. One reason for this was that the state requirements were considerably more specific than were the federal ones. In the study of state programs cited earlier, it was noted that by 1976 seventeen states defined bilingual education as "transitional" — a temporary bridge to help students progress into an all-English curriculum. Thus, it became crucial to develop testing procedures for placing students in bilingual programs, for measuring students' conceptual growth while in those programs, and for assessing English language proficiency to determine when students could move into monolingual (English) classes. Thirteen states had bilingual certification requirements for personnel teaching in these programs. As a result, teacher preparation institutions and state certification boards were put to task to determine what specific knowledge, skills, characteristics, and competencies a bilingual teacher needed. Thirteen states included in their bilingual programs a cultural component recognizing both the importance of self-concept and self-esteem in scholastic success and the need for schools to be sensitive to cultural differences in student behavior as well as in learning styles. Eleven states required strong parental involvement, stressing the importance of the home environment as a part of the total educational experience and the need for the school to understand the sociocultural context in which students are raised. Thirteen states appropriated funds to implement programs. This brought about the development of a variety of program models that were appropriate to the particular needs and characteristics of the population to be served.

Lau v. *Nichols*: a landmark decision

In January 1974, the U.S. Supreme Court decision in *Lau* v. *Nichols* brought national attention to the educational needs of students of limited English-speaking ability. In this case, Chinese public school students claimed that the San Francisco Unified School District was not providing them with equal educational opportunity. The court ruled in favor of the plaintiffs, stating that the district's failure to provide

programs to meet the linguistic needs of the students violated Title VI of the Civil Rights Act. Adding that equal educational opportunity goes beyond providing the same buildings, books, or teachers, it maintained that because these students could not understand the language of the classroom, they were, in effect, deprived of a minimally adequate education (Teitelbaum and Hiller, 1977).

Though not expressly endorsing bilingual education, the *Lau* decision legitimized and gave impetus to the movement for equal educational opportunity for students of limited English proficiency. It brought the needs of those students to the attention of every district receiving federal aid. It set in motion efforts to provide federal enforcement, as well as technical assistance, through a network of regional centers. And it raised the public consciousness of the need for bilingual education, thus aiding the passage of state mandates. The *Lau* ruling also raised many questions that have become very familiar in bilingual education. How many target students must there be? Must they be concentrated in a few schools? What does limited English-speaking ability mean? Is a student from a linguistic minority who can speak and understand English but who reads below level and underachieves in content areas included in *Lau*? What are appropriate remedies? May schools choose whatever program they feel is adequate? Is bicultural education required? What about school desegregation?

Lau Remedies. Following the supreme court decision, the U.S. Office of Civil Rights asked all school districts receiving federal funds to conduct a language survey to identify students of non-English background; this survey subsequently identified over 300 districts that were not in compliance with *Lau*. The immediate issue was, of course, how to go about getting these districts to comply. A set of guidelines called *Lau Remedies* was developed to provide guidance to school districts in assessing students' language development as well as in determining adequate educational programs for them. After assessing the students' home or primary language, the districts were required to assess each student's degree of linguistic function or ability and place him or her in one of five categories (see DeAvila and Duncan, 1976):

A. Monolingual speaker of the language other than English (speaks the language other than English exclusively)
B. Predominantly speaks the language other than English (speaks mostly the language other than English but speaks some English)
C. Bilingual (speaks both the language other than English and English with equal ease)
D. Predominantly speaks English (speaks mostly English but some of the language other than English)
E. Monolingual speaker of English (speaks English exclusively)

Of these categories, only A and E are relatively easy to identify; the others present a problem. One is "struck by the loose manner in which these levels are defined. As such, they bear no resemblance to the 'operational definitions' . . . given in terms of concrete operations, such as scores on tests, numbers of items passed on and so on" (DeAvila and Duncan, 1976, p. 247). For example, category C can really cause problems. The term *bilingual* can be defined in many ways: native-like control of two languages, ability to use two languages alternately, possession of at least *one* of the four basic skills—understanding, speaking, reading, writing—in two languages, and so on. In fact, according to linguists, there are many kinds of bilingualism. Bilinguals are often referred to as balanced (or unbalanced?), coordinate or compound, natural or artificial, bilingual or pseudo-lingual, depending on either how they acquired the languages or how well they command them. One could assume that those who fall into category C are a homogeneous group with native-like proficiency in both languages; in reality, however, a child limited in both English and his or her native language could very well fit into this category since he or she would speak both languages with equal ease (or difficulty). Categories B and D are extremely vague. Since no official definition was offered for *predominantly speaks*, it was left up to the districts to decide.

In some states with mandates, similar but somewhat more explicit categories had been developed for identifying students requiring bilingual instruction. In Illinois, for example, the levels of language fluency were defined as follows (Illinois Office of Education, 1976):

1. The student does not speak, understand, or write English, but may know a few isolated words or expressions.
2. The student understands simple sentences in English, except isolated words or expressions.
3. The student speaks and understands English with hesitancy and difficulty. With effort and help, the student can carry on a conversation in English, understand at least parts of lessons, and follow simple directions.
4. The student speaks and understands English without apparent difficulty but displays low achievement, indicating some language or cultural interference with learning.
5. The student speaks and understands both English and the home language without difficulty and displays normal academic achievement for grade level.
6. The student (of non-English background) either predominantly or exclusively speaks English.

Whereas the *Lau* categories emphasize language dominance, these describe students in terms of English language skills and proficiency as

observed in a school setting, as well as academic achievement. However, this, too, only serves to determine whether or not a student needs bilingual instruction. Once that is determined, assessment of the student's proficiency in the native language is essential in order to prescribe appropriate instruction via the mother tongue.

Resultant Developments. The urgency of complying with the Lau requirements has led some districts to develop useful assessment instruments and procedures. Chicago's Functional Language Survey, for example, includes fifteen items designed to assess the ability of the linguistic minority students identified through the state-mandated census to use the English language. The first five items test the student's ability to repeat sentences said by the rater at normal conversational speed (for example: "I often play with my friends by the fence."). Students are scored on a five-point scale for each item according to accuracy, completeness, and promptness of response. The next five items assess the students' comprehension and elicit verbal responses. (These items might include, for example, "Tell me how to play your favorite game.") The students are again rated on a five-point scale, this time on the basis of comprehension, meaningfulness of response, sentence structure, elaboration, and vocabulary. The last five items do not require testing but are based on students' past performances. The rater is asked to indicate how a particular student would perform five tasks (such as repeating the class homework assignment to English monolingual peers who were not present when it was given). The rater's answer is to be based on the student's oral language performance on the previous test or in school during the past year. After adding all the raw scores for these fifteen items, each student is categorized as Level I, II, III, and so on, according to his or her total score and his or her age.

The San Diego Observation Assessment Instrument, which was also developed to comply with *Lau* requirements, was recently adopted by the state of California to satisfy the requirements of the Chacon-Moscone Bilingual Education Act, AB 1329 (Cornejo and Nadeau, 1978). It is made up of (1) a home language survey; (2) a language observation assessment; and (3) a final assessment. The home language survey consists of four questions (in English and in the home language) addressed to the parents to determine which language a student learned first as an infant, which language the student presently uses in the home, which language adults use in the home, and which language the parents use more frequently with their children. The language observation assessment consists of an interview conducted by a trained bilingual in which a student chooses from a set of "action" pictures and answers a series of open questions asking him or her to list objects in the picture, tell what is taking place in the picture, and expand conversa-

tionally on the picture. (Such questions might include, "What does this make you think of?") On the basis of the responses given, the student is scored at Level I (lists), Level II (tells about), and Level III (expands). In addition, each level is further scored as C = Comprehension, MP = Minimal Production, FP = Full Production, and P = Production at Level III. The final assessment represents the composite estimate of proficiency in the home language and the degree of English fluency demonstrated during the interview. Students are then placed in one of the following categories (Cornejo and Nadeau, 1978):

1. Non-English speaking	*Lau* classification A
2. Limited English-speaking	*Lau* classification B
3. Bilingual	*Lau* classification C
4. Limited other language	*Lau* classification D
5. English only	*Lau* classification E
6. Mixes languages in both interviews	Special
7. No response in either language	Special

Only students placed in categories 1, 2, and 6 qualify for bilingual programs. However, secondary students falling in the bilingual category but scoring below a district's predominant percentile are reclassified as limited English-speaking (LES). Students classified as "limited other language" and "bilingual" also qualify for bilingual instruction if their scholastic achievement is low.

The Chicago and San Diego Language Assessment Instruments, as well as the New York Language Assessment Battery (which responded to the mandate of ASPIRA Consent Decree of 1974 for improved assessment of effectiveness in English and in Spanish (Tilis, Weiciess, and Cumbo, 1978), are designed only for determining whether or not students should be placed in bilingual programs. These are administrative tests developed in response to legal mandates. Their purpose "suits administrative needs rather than pedagogical ones" (Shuy, 1978, p. 376). They help determine the number of students that belong in a given program, but they offer "no hint as to what to do about teaching them." No wonder teachers complain. Needless to say, language assessment for placement is just the tip of the iceberg. Still needed are language proficiency measures for determining treatment procedures to be used in the program. There is also a need to determine what really matters in terms of language proficiency—the more quantifiable and testable features (such as pronunciation, vocabulary, and grammar) or those that are less qualifiable and testable (such as semantic meaning and functional meaning). Shuy argues that functional use of the language is more critical for effective participation than is knowledge of

the language forms in themselves. A student's ability to seek clarification from the teacher for some item is far more important to his learning than are native-like pronunciation and grammar. More must be learned about cultural differences and how they affect learning styles—and hence about needed teaching approaches (Cazden and Leggett, 1976). Appropriate bilingual-program models, as well as instructional and testing materials for diverse groups and circumstances, need to be further developed.

amendments

The amendments to the Bilingual Education Act (ESEA Title VII) made in 1974 addressed many of the needs identified during the implementation of the initial legislation; among other things, they sought a definition of bilingual education, development of bilingual teacher-training programs at the university level, and preparation programs for bilingual paraprofessionals, administrators, counselors, and other support personnel (Schneider, 1976). Greater stress was placed on capacity building, or "a strategy to provide local school districts with the human and material resources needed to operate bilingual programs" (Molina, 1978, p. 23). Since 1974, hundreds of colleges and universities across the country have begun preparing bilingual teachers, the number of graduate programs at the master's and doctoral level have multiplied, a network of support service centers—training resource centers, materials development centers, and dissemination and assessment centers—has been established to help train classroom personnel, provide them with needed curriculum materials, and assist them with all aspects of implementing bilingual education programs. In an effort to help with coordination and to provide technical assistance, funds were allocated for departments of education in the states in which Title VII programs operate. The need for research in bilingual education was also finally addressed; for the first time, substantial funds were allocated for this purpose, as well as for the establishment of a National Clearinghouse for Bilingual Education to collect, analyze, and disseminate information about bilingual programs.

conclusion

These recent efforts in capacity building are beginning to yield results. The expertise and professional preparation of bilingual education personnel have changed greatly from the gut-feeling, common sense approaches of the early seventies. The increasing understanding of the complexities of first- and second-language acquisition and their

implications for diagnosis, placement, and treatment of students is beginning to yield better instruments, instructional approaches, and materials. And the growing body of highly trained researchers specializing in bilingual education is beginning to provide meaningful and responsible studies and evaluations and thus to counteract the effects of incomplete and improperly conducted attempts in the past. In short, we've come a long way in assessing the educational needs of students for whom English is a second language. We have an even longer way to go.

references

Andersson, T., and Boyer, M. *Bilingual Schooling in the United States.* (2nd ed.) Austin, Tex.: National Educational Laboratory Publishers, 1978.

Cazden, C. B., and Leggett, E. L. "Culturally Responsive Education: A Discussion of *Lau* Guidelines, Section II." In *Proceedings of National Conference on Research and Policy Implications, Lau Task Force Report.* Austin, Tex.: Southwest Educational Development Laboratory, 1976.

Cornejo, R., and Nadeau, A. *The California Language Census Survey: Field Methodology Issues.* ICP Occasional Paper Number One. San Diego, Calif.: Institute for Cultural Pluralism, School of Education, San Diego State University, 1978.

DeAvila, E., and Duncan, S. E. "A Few Thoughts About Language Assessment: The *Lau* Decision Reconsidered." In *Proceedings of National Conference on Research and Policy Implications, Lau Task Force Report.* Austin, Tex.: Southwest Educational Development Laboratory, 1976.

DeAvila, E. A., and Havassy, B. "The Testing of Minority Children: A Piagetian Approach." In H. LaFontaine and others (Eds.), *Bilingual Education.* Wayne, N.J.: Avery Publishing Group, 1978.

Development Associates. *A Study of State Programs in Bilingual Education.* Final Report to OPBE USOE. Washington, D.C.: Development Associates, 1977.

Dissemination Center for Bilingual Bicultural Education. *Evaluation Instruments for Bilingual Education: An Annotated Bibliography.* Austin, Tex.: Dissemination Center for Bilingual Bicultural Education, 1975.

Illinois Office of Education, Bilingual Education Section. *Rules and Regulations for Transitional Bilingual Education.* Chicago: Illinois Office of Education, Bilingual Education Section, 1976.

Leibowitz, A. H. "Language Policy in the United States." In H. LaFontaine and others (Eds.), *Bilingual Education.* Wayne, N.J.: Avery Publishing Group, 1978.

Molina, J. D. "National Policy on Bilingual Education: An Historical View of the Federal Role." In H. LaFontaine and others (Eds.), *Bilingual Education.* Wayne, N.J.: Avery Publishing Group, 1978.

Schneider, S. C. *Revolution, Reaction, or Reform: The 1974 Bilingual Education Act.* New York: Las Americas, 1976.

Shuy, R. W. "Problems in Assessing Language Ability in Bilingual Education Programs." In H. LaFontaine and others (Eds.), *Bilingual Education.* Wayne, N.J.: Avery Publishing Group, 1978.

Swanson, M. M. "Bilingual Education: The National Perspective." In G. Jarvis (Ed.), *Responding to New Realities.* Vol. 5: *ACTFL Review of Foreign Language Education.* Skokie, Ill.: National Textbook, 1974.

Teitelbaum, H., and Hiller, R. J. "The Legal Perspective." In M. M. Mammerborn (Ed.), *Bilingual Education: Current Perspective*. Vol. 3. Arlington, Va.: Center for Applied Linguistics, 1977.

Tilis, H. S., Weiciess, W., and Cumbo, R. "On Language Testing: The Development of the Language Assessment Battery." In H. LaFontaine and others (Eds.), *Bilingual Education*. Wayne, N.J.: Avery Publishing Group, 1978.

U.S. Office of Education. *Programs Under Bilingual Education Act: Manual for Project Applicants and Grantees*. Washington, D.C.: U.S. Government Printing Office, 1971.

Zirkel, P. A. "Evaluation and Testing in Bilingual Programs." In H. LaFontaine and others (Eds.), *Bilingual Education*. Wayne, N.J.: Avery Publishing Group, 1978.

Maria Medina Swanson is director of the Bilingual Education Service Center in Arlington Heights, Illinois.

Certain ways of using tests as an element in allocating educational funds are gaining substantial acceptance.

testing and funding: a policy context

joel s. berke

The 1970s have been a time of great change in the way America finances its public schools, particularly at the state level (see Berke and Moskowitz, 1977). One aspect of this change has been a greater role for the state in school finance, a development that has been advocated since the turn of the century, beginning with the work of Elwood Cubberly. The stimulus in the seventies came primarily from judicial interpretations of state constitutions; these judicial decisions required states to change their finance mechanisms to provide greater equity, greater equality, greater equality of opportunity, or more thorough and efficient education, depending on the particular state clause being interpreted. As a result, changes occurred both in the way states and local districts raise revenues for education and in the ways they distribute those revenues. The issue of raising revenue for education can be dealt with briefly, at least as far as tests are concerned, because tests have not been employed in raising revenues for education. However, the fact that revenues for education vary among local districts in each state in direct relation to the availability of taxable property remains an important issue. Thus, the central problem on the revenue side boils down to how to break the link between the availability of taxable property and the amount of money that a local community has for its schools.

One approach to solving this problem assumes that all districts that have chosen the same tax rate will receive equal funding. This outcome may be accomplished by establishing a guaranteed tax base program, by equalizing district power, or by employing various other technical approaches. Most recently, and partly as a response to Proposition 13, we are finding even more interest in systems that ensure greater *parental* choice. Vouchers are under discussion again, and there has been much debate about tax credits at the federal level.

The other major kind of change in school funding involves distributing revenues. Two related issues in this regard are being addressed at the present time in states throughout the country, as well as at the national level. One issue is how to assure that the resources devoted to a child's education do not vary according to where he or she happens to live within a state. How can we break the tie on the spending side between the district a youngster lives in and how much is spent on his or her education? State efforts to solve this problem have led to systems designed to cut down on disparities in spending among districts. Some states have attempted to bring up low-spending districts to a higher level by enhancing the state funding guarantee level, or by increasing the share of funding provided by the state. But there is also a second issue on the spending side of the ledger: how can we ensure a better match between a youngster's need for educational resources and the resources provided to him or her? It is this issue that brings tests into the picture.

testing and federal resource allocations

The federal government, I think, is primarily responsible for building this sort of concern into state funding systems. In the 1960s, the focus of federal aid shifted toward equality of opportunity in an attempt to overcome the disadvantages that some youngsters brought with them when they came to school. Title I of the Elementary and Secondary Education Act is probably the most prominent effort in this direction. This program uses the number of children in poverty as the chief determinant of state, local district, and school attendance area allocations. To determine which children in a target school (a school eligible to receive Title I funds) will actually participate in the program, the criterion shifts to educational need. Tests, as well as other measures, are used to determine which particular pupils will benefit most from the school's Title I allocation.

Since the early 1970s, however, Congressman Quie has led an effort to move the test component of resource allocation (in other words, the educational need component) upward in the allocation chain;

thus, in addition to its role in program participation decisions, testing would be used in designating target schools and in redetermining allocations for the district and the state. Over the last three years, the National Institute of Education (NIE), following a mandate by Congress, studied the feasibility of this approach. But it is clear that any determination of whether such a shift should be made is essentially a value judgment. That is, the NIE study could deal with technical questions, it could estimate costs, and it could even run some experiments in permitting school districts to use tests for fund allocation to schools. It could not, however, resolve the basic question of whether or not it was appropriate to shift from providing funds on the basis of the number of poor children having high educational needs, as at present, to providing funds to meet the needs of all low-achieving children, regardless of their income levels. The basic intellectual determinant in the decision to stay with the poverty criterion for Title I was the recognition that poverty brings problems of its own that deserve special educational treatment and that children living in poverty areas who are having difficulty in school have a harder row to hoe than do middle-class or higher socioeconomic status students who are having difficulty in school. The decision to stay with the poverty determinant has been a value decision, although political determinants have also served to prevent any major changes in the Title I funding formula.

testing and state resource allocations

The 1978 Educational Amendments permit school districts, in certain circumstances, to pick their Title I target schools on the basis of the proportion of children in poverty. However, although the formula for allocating aid among states and school districts is still geared to the number of poor children within those jurisdictions, school districts may now employ tests as well as poverty as criteria for selecting schools to receive Title I funds and (without reference to their parents' income levels) for choosing pupils to receive compensatory services paid for with Title I funds.

The studies conducted by the NIE showed that using tests for the allocation of funds to states and school districts would require new and costly test development efforts. In the thirteen districts that were prompted to use tests experimentally in identifying Title I target schools (under a 1974 congressional mandate to NIE), no radical shifts occurred in the clientele or in the operation of the program. The NIE studies showed that funds appeared to be allocated to a higher proportion of the district's pupils when tests were included as criteria, and, if anything, the concentration of minority pupils receiving Title I services

increased, while income levels rose somewhat. But the trends were not strong, and probably the major conclusion of this aspect of the Title I study was that the test approach to identifying target schools was feasible within districts and caused no major program transformations (see National Institute of Education, 1977a and 1977b).

At the state level, there has been great interest in the last four or five years in trying to relate funding to educational needs. One approach is to have weightings in the general equalization formula for distributing state aid; that is, pupils with identifiable needs, such as the handicapped, are given additional weight. Another approach is to have categorical aid (separate funding for high-need pupils) in addition to equalization aid. In both cases it is possible to identify pockets of educational need through test performance, and some states have adopted such mechanisms. The state of New York, for example, has an extra weighting of .25 for every pupil scoring roughly in the bottom 23 or 24 percent of the pupils taking the Pupil Evaluation Program (PEP) tests at designated grade levels. The test scores are used as indicators of need; they are used to pick out schools and districts with higher than average educational need in order to allocate additional funds in proportion to the number of high-need pupils (see Goettel, 1977). Michigan has had a program for a number of years now in which compensatory education funds are likewise distributed on the basis of test scores; this program has been the subject of much recent discussion (see Murphy and Cohen, 1974). And California has a school improvement act that uses tests to identify pockets of need.

Another way in which states are now using tests as part of their allocation approach is as an accountability measure. For example, the New Jersey supreme court has interpreted a constitutional phrase—the provision of a "thorough and efficient education"—as requiring New Jersey to set educational standards and then ensure that local districts meet those standards by providing appropriate educational treatment for each youngster. Thus, tests are now being used to determine whether the state's responsibility to ensure each youngster a thorough and efficient education related to his or her particular need is being met. This process has also brought about considerable controversy.

conclusion

The use of tests to allocate resources has been under investigation at the federal level since the early 1970s. Tests have been rejected as criteria for distributing federal aid to states and school districts, but they are used to select participating pupils for Title I programs and, as of next year, to choose target schools. In addition, regarding the reform

of state systems of revenue distribution for public elementary and secondary schools, a number of states have sought to attain a direct measure of educational need and have turned to testing as a way either of identifying those areas in need of increased educational services or of showing whether or not the funding system achieves a suitable match between educational needs and available resources.

references

Berke, J. S., and Moskowitz, J. H. "Research on School Finance: A Concern with Equity and Equality." In L. Shulman (Ed.), *1976 Review of Educational Research.* Itasca, Ill.: Peacock Publishers, 1977.

Goettel, R. J. "Educational Needs and State Aid in New York." Paper prepared for section 842 of PL 93-380 project. Albany: State Education Department, State University of New York, October 1977.

Murphy, J. T., and Cohen, D. K. "Accountability in Education: The Michigan Experience." *The Public Interest,* 1974, *36,* 53-81.

National Institute of Education. *Demonstration Studies of Funds Allocation Within Districts.* Washington, D.C.: National Institute of Education, Department of Health, Education, and Welfare, 1977a.

National Institute of Education. *Using Achievement Test Scores to Allocate Title I Funds.* Washington, D.C.: National Institute of Education, Department of Health, Education, and Welfare, 1977b.

Joel S. Berke is a senior research scientist at the Educational Policy Research Institute of Educational Testing Service.

Test scores and determinations of socioeconomic status are used together in an effort to provide an equitable compensatory education program in New Jersey.

testing and funding: the New Jersey experience

fred e. burke

In recent years, the emergence of several educational, social, legal, and economic factors has created a renewed interest in educational funding systems in which student performance on standardized tests determines the amount of money allocated to local school districts. One of those factors is the concern of the public and the educational community over what they perceive as a decline in student test scores. Without accurate measures of the cognitive and affective variables that influence student achievement, it is neither valid nor fair to put great faith in comparisons between past and current student performances on standardized tests; nevertheless, some policy makers assume that tests can identify and define problem areas and that more money will provide the solution to them once they are spelled out. There is an inherent danger of oversimplification on both these counts.

A second factor contributing to the renewed interest in funding systems has been the far-reaching movement of the late 1960s and 1970s to reform school finance. In many states, the courts have provided the impetus for this movement; they have forced legislatures to reexamine the fundamental moral and legal obligations of state governments to provide public school students with a thorough and

equitable education. Today, in contrast, we are experiencing a backlash against state spending in many parts of the country; as evidenced by California's Proposition 13, taxpayers are becoming increasingly unwilling to shoulder growing tax burdens. In fact, tax revolt has spread so quickly that twenty states had anti-tax referendums on their ballots in the last election. Chiefly to blame for this phenomenon is the public sector's failure to maintain high or even mediocre levels of accountability for funded programs; policy makers must now search for ways to provide both an equitable distribution of funds *and* a reliable accountability standard.

The policy behind every funding formula is to distribute dollars in some equitable fashion. All too often, however, we become overly concerned with the funding mechanism and neglect the actual purpose for providing the funds. Therefore, we must decide rather early exactly for whom and for what purpose the money is intended. These are paramount decisions both educationally and socially. For example, should funds go to youngsters who are educationally, culturally, or socially deprived or to children from families with low incomes? These are the kinds of policy decisions that underlie any method that relates funding to test results.

dangers of test-based funding

The concept of test-based funding has changed the definition of equality from emphasizing equal opportunity to stressing equal outcomes. In a few cases, traditional socioeconomic funding models that provided money to the economically deprived have been replaced by test-based funding models; monies are now being distributed to districts whose students perform poorly on standardized tests rather than strictly to low-income districts. We must be aware of this radical change and of the precarious position in which it places educators. We may be offering the schools tremendous disincentives by providing payment for poor, rather than good, student performance.

Under a program of test-based funding, the poorer the performance, the more money a district receives—which is contrary to fundamental educational goals and certainly tempts people to manipulate test scores to gain more money. And, even worse, once funded students manage to reach a predetermined test score, the funding is cut off. There is a real need, particularly among disadvantaged children, for a continuous flow of money if we are to prevent the kind of cognitive and academic regression that followed the withdrawal of funds from various Head Start and Follow Through programs. Still another drawback to reliance on test scores is the concern of many educators—particu-

larly those on the local level—that scores on standardized tests may be misused to evaluate their teaching performance without appropriate consideration of other factors which affect the scores.

The greatest advantage and the most appealing feature of test-based funding is the fact that money is channeled directly to measured student needs. This is an alluring concept, yet we must be wary of its innate pitfalls. For example, the Minimum Basic Skills Test currently employed in New Jersey has a narrow range of content. It focuses, as its title implies, on measuring students' proficiency in reading and computation, the skills considered necessary for minimal functioning in our society today. However, the skills considered basic today may become obsolete in the near future. We must concern ourselves with a broader range of programs and curriculums, particularly for helping our disadvantaged children. Such children may master the basic skills and not require any additional money under a test-based funding scheme, yet they may be socially and emotionally below society's standards and desperately need exposure to the socializing and maturing aspects of a compensatory education program. Social institutions in this country are filled with people who possess the basic skills but lack either the ability or the desire to progress beyond the welfare or unemployment line or, even worse, prison.

alternative funding approaches

Alternatives to test-based funding, such as various equalization formulas, are currently used in many states. Equalization formulas are designed to reduce disparities in per-pupil expenditures between districts. Equalization is easy to administer, but equity is hard to achieve. For example, we have almost equalized the local district tax effort in New Jersey but we still cannot generate the same amount of financial support for every public school student. We allowed local decision makers to determine where they wanted these funds to go, and, for reasons of political expediency, educational expenditures were not equalized. We can equalize taxing capacity but we cannot equalize the values that are placed on education or determine the priorities that people place on their actual or perceived needs. In short, equalization formulas do not necessarily benefit needy students, regardless of how they are funded.

The allocation of Title I funds is based on socioeconomic indicators such as census data, Aid to Families with Dependent Children (AFDC) counts, parental income, and parental education. Obviously, socioeconomic variables do not necessarily address the needs of or even identify all students who might need compensatory education. Also,

socioeconomic data, particularly census data, are often outdated by the time they are compiled. Nor do these indicators compensate for rapid population changes or shifts, a fact that has been a major issue since the advent of declining enrollment and inter- and intradistrict student transfers. In addition, local surveys intended to measure socioeconomic status are often difficult to develop, have questionable reliability, and are sometimes an intrusion on people's privacy. The AFDC count, which is updated annually, is an acceptable measure, but it is still a substitute. Unfortunately, socioeconomic indicators provide us with information about corollary types of associations that all too often do not accurately distinguish needy children from others. Lacking any more useful information, we make assumptions that have to do with parental education and parental income because they are categories that we can measure. In this way, we often confuse the purpose with the cause and begin to address problems (especially those of inner cities, such as poverty) that are beyond the range of education. In order to be effective, the Title I funding approach must be comprehensive, well-organized, and aimed at providing long-range answers to the problems of social, cultural, and economic deprivation. In essence, the short-term, narrow-range funding approach can provide only short-term, narrow results. Our society's ship will continue to sink if we continue to plug only one of the many holes in our hull.

learning from experience

New Jersey has a unique compensatory education funding formula. We use socioeconomic *and* testing criteria to determine the levels of compensatory aid that will be allocated to local districts. Our systems relies, in part, on a statewide assessment of academic achievement. This test is administered annually to identify children whose scores fall below a state minimum standard of performance in grades three, six, nine, and eleven; we can then estimate the scores of all grade levels that were not tested. However, we believe that an approach based entirely on test results would not serve a significant number of the children in need of compensatory education. A test can provide a direct measure of needs, but it fails to address the causative factors related to low achievement and, therefore, might lead to neglect in areas that we do not test. We thus allocate funds on the basis of two indexes—test scores and socioeconomic status as indicated by AFDC count. In any case, our law makes the decision very clear; it states that funds must be addressed to those who are educationally, socially, culturally, and economically deprived.

I believe that the system in New Jersey makes sensible use of

both socioeconomic indicators and test results, which enhances our ability to provide an equitable funding program. We give a lesser weight to the test measure than to the socioeconomic measure and thus both minimize the disincentive for high achievement characteristic of simpler and purely test-based approaches and discourage core manipulation. Moreover, we can be relatively up-to-date because AFDC information is updated annually and we test annually. However, while this system seems to have worked well so far, we realize that we need to provide an incentive to students to improve their scores. Some solutions we have considered are maintaining the same funding level over a three-year period or providing an initial base level of funds and then allocating additional money after improvement is demonstrated.

reluctant recognition

Many New Jersey legislators are not enthusiastic about test-based funding. One reason for this is its rising cost. When it was begun in 1976, the New Jersey State Compensatory Education Program cost $32.8 million; this year (1978), the cost is up to $68 million. This situation presents us with a dilemma. If the need for compensatory education lessens, the funds will dry up. If the need becomes greater, there will be greater reluctance to fund the program. In either case, some children may be left without the compensatory education they require.

Many legislators today are becoming increasingly concerned that one third of the entire state budget is to be spent on education. What is more, they are concerned that, with our formulas, the decisions have already been made and all they can do is vote for or against them. It is perhaps understandable that the legislators seeking to make the fundamental decisions want to recenter the decision-making process. In addition, legislators are becoming increasingly sensitive to the attitudes of our older citizens. As the voting population becomes more and more one of people who do not have children in public schools, we are going to find a broadening credibility gap and greater dissatisfaction with our political processes and the way in which we determine the amount of money available for education. Funding allocations will no longer be made by means of what we know as the normal democratic process as the problem of semi-unlimited public programs competing for finite resources continues to grow. This is something which is going to happen throughout the country. In New Jersey, I should add, we are legally mandated to give account to the legislature on the status and adequacy of the use of the formula in allocating funds.

Finally, the legislature and state board of education are contemplating the implementation of high school graduation standards

that will involve testing to determine eligibility not only for receiving diplomas but for remedial programs as well. One question I think we have to ask ourselves is to what extent such an additional remedial program is test-based and thus will require additional funds. Where will those funds come from and to what extent will existing compensatory dollars have to be reallocated from the lower to the higher grades? These questions have implications for every state.

Statewide achievement tests were intended to identify children in academic difficulty. They provide information that is needed at the classroom and individual school levels. Unfortunately, however, that information is all too often put to the wrong use. In New Jersey, for example, results of statewide achievement tests are one consideration in evaluating tenured teachers. The test results are only one of several factors considered in those evaluations, but their use in this way has created and will create anxiety among teachers about their continuing employment prospects.

upcoming policy decisions

Are all these uses of state test data compatible? Is it logical, for example, to use the state test both to evaluate schools and to distribute compensatory dollars to pay for the discrepancies identified by such evaluations? In New Jersey, we must perform a district evaluation of every school each year. Beginning next year, we will have to classify every school as approved, unapproved, or conditionally approved. A critical component in this classification is how well students do on tests. In my opinion, we are asking one less-than-perfect instrument—the standardized test—to bear too much responsibility in educational decision making. Would not too many uses of test data exert excessive influence in the allocation of educational services?

A second set of policy questions concerns whether or not test-based funding can or even should survive in a period of fiscal contraction. How much of our educational resources should we distribute through a test mechanism? What should be the role of the people's representatives in a democratic system—should *they* not be the ones who determine the proportion of funds that should be distributed rather than leaving it to a formula? What if the students' scores do not improve? What happens then? Does this not provide extraordinary ammunition for the decision makers who are no longer child advocates? They may say that test scores show that dollars do not make any difference. If and when that happens, scores will become extremely dangerous for those of us who try to get the maximum amount of money into public education. There is no direct correlation between

the amount of money we invest and the scores that emerge as a consequence of that investment.

Is there a limit to the funds that are needed for a test-based fundng mechanism? The legislators say: "Commissioner, your program started at $30 million, and in three years you have increased it to $68 million. When is it going to stop? How do you get the children out? You now have a program that is bigger than Title I." Since our total available resources are growing at a slower rate than are our test-based allocations, test-based funding can only siphon away from other programs. We have already seen this beginning to happen.

At some point, we must decide whether consistently low test results warrant new money for the same old programs or if we should look to more radical alternatives. How do we decide when we have reached such a critical point? And what will happen to test-directed funds when test scores improve, as indeed they must if our current remedial efforts have any validity? These are the kinds of questions we are beginning to ask and must try to answer not only in New Jersey but throughout the United States.

I believe that increased reliance on testing in educational decision making, particularly when tests are used for allocating funds, may create difficult policy problems. This approach might inadvertently lead to a reduction in resources for public education. This approach tends to over-formalize and over-simplify allocation decisions and, thus, to leave out key steps in the decision process.

In New Jersey, we have found that combining test data with measures of poverty gives us a balanced system that maintains the testing component in its proper role. Nevertheless, we realize that funding mechanisms must be subjected to continual scrutiny to ensure that they are achieving their purpose.

reference

Mathis, W. J. "The Use of Basic Skills and Socioeconomic Data in Determining State Compensatory Funding Entitlements." Paper presented at the American Educational Research Association Annual Meeting, Toronto, March 1978.

Fred E. Burke is Commissioner of Education of the New Jersey Department of Education.

Broad educational consequences of test-based funding must be taken into account in designing and evaluating funding plans.

testing and funding: measurement and policy issues

george f. madaus

The marriage between funding and test performance was first proposed by a select committee of the Irish Parliament in 1799; sixty years later, the match was finally arranged by Robert Lowe. In a time of severe strain on the exchequer brought about by the Crimean War, and in a time of increasing enrollment and concern over standards, Lowe tied the knot by formalizing the following recommendation of the Newcastle Commission (Coulahan, 1975, p. 75):

> A searching examination . . . should be made . . . of every child in every school . . . with the view to ascertaining whether these indispensable elements of knowledge are thoroughly acquired and to make the prospects and positions of the teachers dependent to a considerable extent on the results of this examination.

The match became known as payment by results. For better or for worse, it was predicated on the assumption that there is a positive incentive in linking teachers' salaries to pupil achievement on written and oral examinations in reading, writing, and arithmetic. Over the next three decades, the compatibility of testing and funding was

severely strained and the two separated in England at the turn of the century. Today, once again, in a time of rising expenditures, accountability, and increased concern about standards, proposals such as that by Congressman Quie and programs such as Michigan's Chapter Three have reunited testing and funding. Now, however, testing is used to indicate where funds for compensatory programs or remedial assistance should be allocated.

In order to fully understand the current relationship between testing and funding, several considerations that are central to present proposals for test-based funding need to be recognized. These include: the social implications of shifting the operational definition of educational disadvantage from an index of poverty to one of poor test performance (see Feldmesser, 1975; Kellaghan, 1977); the numerous technical, psychometric, and administrative issues associated with implementing specific proposals (see Feldmesser, 1975; Haertel and others, 1977; Madaus and Elmore, 1973); and related assumptions that the funds actually provide *additional* services to the disadvantaged and that the schools already know how to remedy the deficiencies that lead to low test performance (see Airasian, 1978; Airasian, Madaus, and Pedulla, 1978). These considerations, however, have been treated elsewhere and are beyond the scope of this chapter. Here I would like to consider test-based funding more broadly. I will argue that it is one indicator of an inexorable but unconscious populist movement in many states toward a system of public, or external, examinations. I will describe the mechanisms by which tests can give an external agency various degrees of control over schooling. And, finally, I will evaluate the degree of such control that various proposals for test-based funding would give to the external agency.

new support for external testing

Tests developed by an agency outside the school have commonly been used by governments to certify students' success at one level of education and then admit them to either the next level or to civil service or other careers. A system of external tests, while not unknown in this country (witness the New York Regents Exams and the College Board tests), is, nevertheless, a rather alien concept that is more common to British and other European systems. As unattractive as such a system may be to American educators, however, the public is moving toward acceptance of testing by an agency outside the school, according to recent opinion polls. A 1976 Gallup survey found that 65 percent of the public agreed that pupils should pass a state or national exam in order to graduate from high school. In a more recent (1978) Gallup survey, 68 percent of the general public felt that pupils should

be promoted only if they pass an exam, and 53 percent felt that such an exam should be prepared by either the state or the national government. In my home state of Massachusetts, a recent poll showed that 83 percent of the public favored a basic skills examination as a requirement for high school graduation (Clark University, 1978).

At a time when testing is receiving fierce criticism from many academics, civil rights advocates, and professional organizations, these poll results illustrate an interesting dichotomy of attitudes. While critics are castigating testing, taxpayers, parents, businessmen, legislators, and much of the media are demanding more testing to increase accountability, return to "basics," eliminate the influence of social position, ensure minimal competencies, and improve standards.

The expanding movement toward certifying minimal competency for graduation and the increase in proposals for test-based funding are two conspicuous and explicit indicators of a trend toward external testing. Florida's minimal competency tests, which would link high school graduation to state-level tests, are a clear example of the former (Haney and Madaus, 1978; Madaus and Airasian, 1977). The Quie bill and legislation in Michigan and more recently in Connecticut are proposals for test-based funding. These proposals involve external testing programs because of the need for comparable test data at the state or district level when funds are allocated for remedial assistance.

While the movement supporting external testing programs has been pressing forward relentlessly in the states, it appears to be dead at the federal level. The National Institute of Education (NIE) is on record as being opposed to both national minimal competency tests (Graham, 1978) and federal test-based funding (National Institute of Education, 1977). The reasons for this are not primarily technical or practical, although there are many interesting and complex administrative and methodological problems inherent in test-based funding (see Harnischfeger and Wiley, 1977a, 1977b; Madaus and Elmore, 1978); the reasons are fundamentally political. Powerful educational lobbies have opposed both plans because they correctly perceived that such testing programs could dramatically shift control of the curriculum to the federal level. The same argument, of course, holds true at the state level, but there proponents of test-based funding and minimal competency programs have been much more successful.

the power of proficiency exams

When results of external exams are the sole or even a partial determiner of future educational or life choices, or when they are used as a means to provide positive incentive in a substantial funding scheme, they influence what is taught, how it is taught, what pupils study, and

how they study (see Madaus and Airasian, 1977; Madaus, Kellaghan, and Airasian, 1971; Madaus and Macnamara, 1970; Srinivasan, 1971). The mechanism for such control involves the need to agree on a set of objectives that transcend district boundaries. This in itself is a sticky point for many since these objectives, although perhaps minimal, might, in fact constitute a national, or more likely a state-level, syllabus. However, it is the test that measures this syllabus and that is used to monitor, certify, or allocate funds that is the linchpin of the control mechanism. Control over the curriculum, teaching, and learning is mediated through a process that Europeans call the "tradition of past exams." In most external exam programs—the College Board's exams and Florida's minimal competency program are notable exceptions—the tests move directly into the public domain once they are administered; over a period of time, teachers, pupils, and parents learn to infer from the tests what is important. In reality, the tradition of these tests defines the important objectives of the schools. It is this tradition that gives the testing agency the potential for enormous control over the curriculum and, consequently, over the teaching and learning process.

Such control is a double-edged sword. On the positive side, well-defined and valid performance measures have been powerful forces for redirecting teaching and effecting curricular change (see Bloom, 1950; Commission on Mathematics, 1959; Morris, 1969). Given our present emphasis on a return to basics, or mastering minimal competencies, this could be an important benefit. On the negative side, however, most studies have found that curriculum, instruction, and learning regress to the tradition of the tests; the proportion of instructional and study time spent on various elements of the curriculum is seldom higher than the predicted likelihood of their occurrence on the exam (see Madaus and Macnamara, 1970; Norwood Report, 1943; Spaulding, 1938; Srinivasan, 1971). Further, the Irish Intermediate Board of Education (1971), during the payment by results era, articulated a now familiar complaint when it deplored interschool comparisons "that forced schools into competition with one another—a competition which is naturally injurious to the best interests of secondary education" (pp. xi, xii). Present proposals and programs for test-based funding or for certifying minimal competencies using norm-or criterion-referenced tests certainly permit and encourage interstate or inter-district comparisons (Madaus and Elmore, 1973).

options for test-based funding

The amount of money available and the way it is allocated determine the extent of control exercised by the external agency

through its tests. In the nineteenth century, when test results were a key element in fixing a teacher's salary, the effects on all parties were devastatingly negative (Herbert, 1889; Holmes, 1911). Exacerbating the situation was the fact that the tests used by the school inspectors changed very little from year to year. Matthew Arnold (1899, p. 136) then one of the school inspectors, cynically described the payment by results system as a "game of mechanical contrivance in which the teachers will and must, more and more, learn how to beat us." Rather than teaching for the test, teachers were eventually able to teach the test itself—to cram their pupils with the answers to perennial questions.

In the recent past, we have seen the emergence of performance contracting, a close relative of payment by results. In both cases, money was linked to test gains. In performance contracting, the contracter receives payment; in payment by result the teacher receives payment. Like its ancestor, performance contracting often substitutes cramming for learning. To my knowledge, there has never been an attempt to link substantial financial incentives to a *new* test each year based on a stable but well-defined domain of minimal objectives. If such a system were attempted, the tradition of the tests would soon become a powerful force in the schools. One could predict that, after a few years, the distribution of those passing the tests would stabilize at a very high percentage. If we are talking about basic or minimal skills, some may argue that this is exactly the distribution we want. But there is a tradeoff. Given our testing history, the multiple-choice format might be expected to quickly and uncritically dominate the external test and thus might, unfortunately, influence the kind of teaching and learning that takes place.

An alternative to a positive incentive, test gain funding plan is one that links funding levels for remedial assistance programs to low test performance. This is still an external testing program, but it is one whose effects on the curriculum and on teaching and learning should be slight—so long as safeguards are built in to discourage schools from implicitly or explicitly taking steps to depress scores on which funds are allocated (see Feldmesser, 1975; Harnischfeger and Wiley, 1977b) and so long as the continuation of funding is not linked to test score gains. This describes the current situation in Michigan. However, if continuation funding is reduced when pupils make test gains, as under Quie's plan, a strong negative incentive is introduced. To avoid such a negative incentive, the Michigan Chapter Three legislation originally set up a two-tiered testing program that tested initially to allocate funds for low-scoring pupils and then again to link continuation funding to successful test performance. Districts would receive full allocation the following year for each low-scoring pupil who achieved 75 percent of

agreed-on objectives and a proportionate amount for partial gains. This use of test results for continuation funding caused considerable controversy. It was perceived not as an incentive but as a penalty—a device that would be used to single out teachers with pupil failures (Murphy and Cohen, 1974). Consequently, this continuation funding component of Chapter Three has never been implemented. If it had been, it is likely that the tests eventually would have influenced teaching and learning in Michigan schools.

Mosher (1973) suggests an interesting use of a two-tiered system of test-based funding. He feels that commercial norm-referenced achievement tests are the most suitable devices for initially allocating funds for remedial assistance programs. He suggests, however, that a different type of achievement test is best for evaluating the effectiveness of these programs or for making decisions about continuation funding. It can be argued that, for a number of reasons, norm-referenced achievement tests tend to measure general ability rather than school-specific achievement. Such achievement tests correlate as highly with so-called intelligence or verbal ability tests as they do with one another, and they also correlate highly with home background (Coleman and others, 1966). Thus, they afford a realistic index of the difficulty the school will have in teaching low-scoring pupils. However, because of their psychometric properties and the collusive effect of home and school on the traits they measure, these general achievement tests are not particularly sensitive instruments for assessing changes in the school's effectiveness in reaching specific instructional objectives. Thus, Mosher (1973) argues that they should not be used to evaluate the effectiveness of programs. Instead, he suggests that tests geared specifically to programs' instructional objectives be employed.

The effects of using such objective-referenced tests to evaluate programs should be benign if the funding level determined by the norm-referenced achievement/abilty tests is not affected by the outcome of the evaluation. However, if continuation funding is tied to gains on a test referenced to a set of common statewide objectives, then the potential impact of that test could be great indeed. Thus, I should like to suggest a variation on Mosher's plan. Like Mosher, I would first allocate funds on the basis of a general norm-referenced achievement/ability test. The state could require districts to modify their programs on the basis of subsequent evaluation results but could not use those results to reduce the initial funding level. However, a bonus might be paid to districts for every economically disadvantaged pupil whose scores reach some agreed-on standard on a test geared to a set of competencies for a particular grade. Safeguards would need to be built into the program to avoid the segregation of these bonus eligible students,

to keep them from being shortchanged in other aspects of the curriculum, and to guard against accepting minimums as the norm. Such a plan might capitalize on one positive impact of external exams—the structuring and focusing of instruction—for the students the schools have had the least success in reaching: the economically disadvantaged.

are we ready for government testing?

Whether or not compensatory funds should be allocated on the basis of test scores comes down to deciding whether or not our society is willing to accept a federal (or more likely a state-level) external testing system. The acceptance of such a program would alter the present system of American testing. Instead of a system in which local districts use privately developed tests in traditional ways (which I feel have minimal impact on the schooling process) we would move to a system in which tests used by the state may have a profound influence on the curriculum, as well as on instruction and learning. The effect on the balance of power between the local district and the state would be a direct function of the rewards or sanctions associated with the use of the external tests.

Can a system of test-based funding be built that could alter the present balance? Absolutely! Should we then move in this direction? That is not primarily a measurement question, although there are measurement issues involved; it is a question of values, politics, power, and control. Whatever society decides, we must be aware that a system of external testing linked to funding involves a delicate balance; it is *not* a marriage made in heaven.

references

Airasian, P. W. "Pre-measurement Issues in Minimal Competency Testing Programs." Paper presented at the National Consortium on Testing, Spring Membership Conference, Arlington, Va., May 1978.

Airasian, P. W., Madaus, G. F., and Pedulla, J. J. *Policy Issues in Minimal Competency Testing and a Comparison of Implementation Models.* Commonwealth of Massachusetts, Department of Education. Wellesley, Mass.: Heuristics, 1978.

Arnold, M. *Reports on Elementary Schools, 1852-1882.* (Edited by Sir Francis Sandford.) London: Macmillan, 1889.

Bloom, B. S. *Problem-Solving Processes of College Students.* Chicago: University of Chicago Press, 1950.

Clark University. "Large Majority Support Basic Skills Examination Requirement for High School Graduation." News release, Worcester, Mass., 1978.

Coleman, J., and others. *Equality of Educational Opportunity.* Washington, D.C.: U.S. Office of Education, Department of Health, Education, and Welfare, 1966.

Commission on Mathematics. *Program for College Preparatory Mathematics: Report of the Commission on Mathematics.* New York: College Entrance Examination Board, 1959.

Coulahan, J. M. "The Origins of the Payment by Results Policy in Education and the Experience of It in the National and Intermediate Schools of Ireland." Unpublished master's thesis, Trinity College, Dublin, 1975.

Feldmesser, R. A. *The Use of Test Scores as a Basis for Allocating Educational Resources: A Synthesis and Interpretation of Knowledge and Experience*. Princeton, N.J.: Educational Testing Service, 1975.

Gallup, G. H. "Eighth Annual Gallup Poll of the Public's Attitude Toward the Public Schools." *Phi Delta Kappan*, 1976, *58* (2), 187-201.

Gallup, G. H. "Tenth Annual Gallup Poll of the Public's Attitude Toward the Public Schools." *Phi Delta Kappan*, 1978, *60* (1), 33-45.

Graham, P. A. "Remarks." In W. Haney (Ed.), Summary of the Spring 1978 Conference of the National Consortium on Testing, Washington, D.C., June 5, 1978.

Haertel, E. H., and others. *Achievement Measures as Title I Eligibility Criteria: Concepts, Methods, and Eligibility Estimation*. Chicago: ML-Group for Policy Studies in Education, CEMREL, 1977.

Haney, W., and Madaus, G. F. "Making Sense of the Competency Testing Movement." *Harvard Educational Review*, 1978, *48* (4), 462-484.

Harnischfeger, A., and Wiley, D. E. *Statement to the Subcommittee on Elementary, Secondary, and Vocational Education*. Chicago, Ill.: ML-Group for Policy Studies in Education, CEMREL, 1977a.

Harnischfeger, A., and Wiley, D. E. *A Study of Student Achievement Measures as Title I Eligibility Criteria*. Chicago: ML-Group for Policy Studies in Education, CEMREL, 1977b.

Herbert, A. (Ed.) *The Sacrifice of Education to Examinations*. Edinburgh, London: Williams and Norgate, 1889.

Holmes, E. G. A. *What Is and What Might Be: A Study of Education in General and Elementary in Particular*. London: Constable, 1911.

Irish Intermediate Board of Education. *Report of the Intermediate Education Board for the Year 1911*. (CD 6317) H.C. 1912-1913.

Kellaghan, T. *The Evaluation of an Intervention Programme for Disadvantaged Children*. Berks, Ireland: NFER Publishing, 1977.

Madaus, G. F., and Airasian, P. W. "Issues in Evaluating Student Outcomes in Competency-Based Graduation Programs." *Journal of Research and Development in Education*, 1977, *10* (3), 79-91.

Madaus, G. F., and Elmore, R. F. *Allocation of Federal Compensatory Education Funds on the Basis of Pupil Achievement Test Performance*. Hearings before the General Subcommittee on Education of the Committee on Education and Labor, House of Representatives, 93rd Congress. Washington, D.C.: U.S. Government Printing Office, 1973.

Madaus, G. F., Airasian, P. W., and Kellaghan, T. "The Effects of Standardized Testing." *Irish Journal of Education*, 1971, 2, 70-85.

Madaus, G. F., Kellaghan, T., and Airasian, P. W. "The Stability of Teachers' Perceptions of Pupil Characteristics." *Irish Journal of Education*, 1975. *11* (3).

Madaus, G. F., and Macnamara, J. *Public Examinations: A Study of the Irish Leaving Certificate*. Dublin: Educational Research Centre, St. Patrick's College, 1970.

Morris, G. C. "Educational Objectives of Higher Secondary School Science." Unpublished doctoral dissertation, University of Sydney, 1969.

Mosher, F. "Memo to Frank Keppel re H. R. 5163 – The Quie Amendment to Title I, ESEA." New York: Carnegie Corporation of New York, 1973.

Murphy, J. T., and Cohen, D. K. "Accountability in Education – the Michigan Experience." *The Public Interest*, 1974, *36*, 53-81.

National Institute of Education. *Using Achievement Test Scores to Allocate Title I Funds*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1977.

Norwood Report. *Curriculum Examinations in Secondary Schools*. Dublin: H. M. Stationery Office, 1943.

Spaulding, F. T. *High School and Life: The Regents' Inquiry into the Character and Cost of Public Education in the State of New York*. New York: McGraw-Hill, 1938.

Srinivasan, J. T. "Annual Terminal Examinations in the Jesuit High Schools of Madras, India." Unpublished doctoral dissertation, Boston College, 1971.

George F. Madaus is professor of education at Boston College.

High school graduation should mean that the school has provided a suitable program of learning activities and that the student has attained defined levels of performance.

tests and diplomas: certifying high school education

mark r. shedd

Tests and diplomas are the subject of considerable discussion these days—both inside and outside of education circles. Their purpose, in many minds, is to certify or validate education. Frankly, I think we could benefit from a careful look at both.

The high school diploma has long been a symbol of student accomplishment, representing the sum total of personal and academic achievement. But recently it has taken on new significance. Now it is expected to certify not only individual attainment but also the success of the schools in providing quality education. This is a subtle but important shift in emphasis. As a mechanism for public accountability, the diploma must have more than personal, individual significance; it must have universal validity. And that, in turn, requires measurable standards for earning a diploma.

This shift in expectations has produced tremendous pressure to move in the direction of standardized tests, as well as raging controversy over the subject. Supporters of minimum competency testing point to illiterate graduates, frustrated parents and employers, social rather than competency-based promotions, and the like as reasons for

using minimal performance tests as standards for high school graduation. But opponents argue that such tests prove little, educate not at all, and can be seriously misused. Minimum expectations for schools, they fear, will soon become maximum expectations for students.

While my own bias is with those opposed to such testing, I believe that both sides are correct—at least insofar as their facts are concerned. There *are* serious problems afflicting the schools, and they demand our immediate attention. However, testing is a simplistic reaction to a complex set of problems that demands a more thoughtful response.

accountability in education

The fundamental issue here is *accountability*. That very popular word represents one of the most basic foundations of our democratic society: the ability of the people to demand that their public institutions account for the quality of their work. This is perhaps particularly true for our educational institutions. Because we Americans value our schools enormously, we expect a great deal from them and we spend a great deal on them; we therefore have every right to demand accountability from the institutions we support.

But what exactly do we expect? Despite arguments that the schools have tried to take on too much, there are broad areas of agreement about the purposes of education. In general, we want the schools to help our children learn to communicate and compute, to become capable of making a living, and to be good parents and neighbors, as well as wise consumers and voters. Schools should help children form and express opinions, make judgments, solve problems, be creative, and enjoy their own lives and the world around them. To borrow from other writers and educators, we want our schools to enable children to: find pleasure in the exercise of their minds, to help them realize their potentialities, to educate themselves throughout their lives.

We demand a great deal. And we are deeply concerned about the quality of education in America. The latest Gallup poll tells us that two thirds of the American public believes the quality of education is declining. And, while I continue to argue that schools today do a better job of educating more youngsters than ever before, I recognize that there *are* students who are not learning; there *are* teachers who are not teaching; and, therefore, students, parents, and taxpayers are being cheated by the schools.

Clearly, we face a difficult dilemma: we must spend our energy addressing the public concern about the quality of our education while

at the same time attempting to analyze and resolve the problems involved. The crux of the matter is that education is hard, if not impossible, to quantify. That is a galling realization for the state legislator whose constituents demand that something be done about the quality of education. It is frustrating for a reporter seeking a neat definition of a "successful" public school. And it leaves no recourse for a parent who suspects that his child is not being educated. So we attempt to "measure" education with competency and proficiency tests. And so great is the pressure for accountability that thirty-three states now use some form of standardized testing and every other state is considering it.

While apparently logical to many people, the testing response has many flaws. It is a simplistic approach that ignores the drawbacks of proficiency and competency tests: they are limited instruments measuring limited numbers of things; they are always biased in some way; and they can identify problems but not causes or solutions. It also ignores the potential for misuse of proficiency tests. If used to deny promotion or graduation—a practice that has never been proven to benefit students—tests have the effect of blaming students for schools' failures. Such tests are equally unsuitable for use as the sole judge of a school's success. I am deeply disturbed by the growing tendency to compare one school with another *only* on the basis of standardized test scores. That "bottom line" approach is a meaningless device of the business world that is unfair to students, schools, and the public that believes it to be valid. There are, of course, many valid uses of tests; these range from diagnosis of individual learning problems to the evaluation of whole programs over time. Furthermore, tests should play an important role in the overall accountability process. But tests are inappropriate as the ultimate measure of education; no test has been proven to accurately predict success in adult life.

an alternative to testing

The demand for accountability—from students, parents, and taxpayers—is legitimate and, practically speaking, too powerful to ignore. We must devise some valid way of certifying that an acceptable process of education occurs between the first and twelfth grades. One reasonable alternative to testing is taking shape in Connecticut. This alternative is not the perfect solution to education's problems; nor is it entirely independent of the testing approach. With its establishment this year of a statewide proficiency exam and with its proposal last year of a statewide competency-based test for granting high school diplomas, Connecticut hopped on the testing bandwagon, too. But there is

some important restraint in this testing program: Connecticut is easing into a comprehensive program of accountability that includes testing but is not limited to it.

The proficiency and competency tests I have just mentioned are important components of the accountability program and warrant some explanation. The proposed competency-based high school diploma test is the more controversial of the two, because even though it is an optional program, aimed mostly at out-of-school youth, it does establish statewide criteria for a high school diploma.

Like most states, Connecticut has had a state testing program for high school equivalency for years. Limited to those over age twenty, this test gives adults the opportunity to earn a high school diploma. However, a year ago a study group, established under Connecticut's Master Plan for Vocational and Career Education, recommended that a measure of "competency" or applied skills be added to the test; the group believed that the high school diploma should reflect such competency as well as demonstrated academic proficiency. The proposal was aimed chiefly at young men and women—many of them dropouts—between the ages of sixteen and twenty who might not otherwise have the opportunity to earn a diploma. The study group also proposed lowering the eligibility age for the test, thereby permitting some students to graduate early.

A source of controversy because of that "early exit" provision, the proposal awaits further action and funding from the General Assembly. However, there are a number of features—details that do not make headlines—of the proposed test which would make it an important part of the overall accountability process. (This is a very important option for some young people who may not have the choice of staying in school for the last year or two. For them, the opportunity to earn a high school diploma and the earning power that goes with it are essential.) Our proposal would allow students to "test out" of school only with parental permission, and then only after intensive counseling. Students who had dropped out of school were also to be given counseling and encouraged to return to class to earn a diploma.

At the time that it was asked to act on this proposal, the Connecticut General Assembly was preoccupied with measures for proficiency testing. Since then, Connecticut has joined the mainstream in passing a proficiency testing bill, although with some restraint. For one thing, the bill is called the Education Evaluation and Remedial Assistance Act, and it means just that: it is based on evaluation and remedial assistance, *not* competency tests and *not* requirements for promotion or for graduation. It will not be the ultimate arbiter of students' success or failure. The act is intended to evaluate student proficiency in basic academic skills and to assign remedial assistance where

needed. Furthermore, it is not just one test. The act requires local districts to test students at three grade levels between the first and eighth grades, and it calls for a fourth test to be administered by the state to all ninth graders. The choice of grades is intended to provide ample time to correct problems indicated by the tests. The law also requires each district to plan a comprehensive testing program before administering it, and it assigns state money for remedial efforts.

As state tests go, this is not a bad model for an accountability device. It is aimed primarily at schools' not students and its chief purpose is to provide aid - not labels - for students. Furthermore, it steers clear of the testing-for-promotion trap. But it is not a perfect solution. I would prefer to see the statewide test administered at both grades four and eight rather than grade nine, thus allowing more time for remedial help. And I am still deeply concerned about the possible misuse of test results. After our experiences with SAT scores, I cannot believe that this will not be a problem. Realtors will find the scores helpful in identifying "good" school systems; parents and taxpayers will use them to compare students, schools, districts, and teachers. And, while I know such use is distorted, unfair, and unhelpful, I also know it is unavoidable -- particularly since state money will be allotted to towns on the basis of the number of students who fail these tests. In developing regulations for the law, we will be working to minimize this problem as much as possible.

The most encouraging aspect of this testing legislation is that it is regarded in a broad context of accountability. By and large, neither the legislature nor the public assumes the test to be the sole answer to the accountability issue. This is primarily because Connecticut is dealing with a larger issue at the moment -- the state supreme court ruling in *Horton* v. *Meskill* that Connecticut's system of financing schools is unconstitutional. Like California, New Jersey, and other states, Connecticut is thus faced with the prospect of redesigning not only the financing of education but the structure of that education as well.

examining "suitable" education

Our statutes not only demand equal opportunity for all students and a reasonable level of funding; they also require that opportunity be provided each student for something called a "suitable program of educational experiences," which, to date, has never been officially defined. In order to shape an equitable finance system, it became clear very early that we would have to define a "suitable" education as well. That process is not yet complete, although a final proposal is now being prepared for review by the school finance advisory panel, the state board of education, and the General Assembly. Nonetheless,

after months of discussion, debate, criticism, and advice, a consensus has emerged that gives shape to a definition of suitable education. So far, we have successfully resisted the urge underlying the competency movement to quantify "suitable" in terms of a specified list of programs or student requirements. There will be no state curriculum or state graduation requirements — and that represents more than just a deferral to New England's penchant for local autonomy. It reflects an understanding of the education process as one that must be flexible, responsive, and individualized, not merely convenient for adults.

Connecticut's "suitable" education program is shaping up as a series of guarantees to students of appropriate opportunities for their education. The guarantees will also assure parents and taxpayers of the accountability they demand for school performance. Among the elements of a suitable education program are various state and local goals and objectives; minimum curricular offerings; minimum funding; appropriate staffing, equipment, and supplies; adequate systems for managing, evaluating, and improving school programs; and, finally, an effective evaluation and reporting system including the various testing programs mentioned earlier. In addition, the key to the process is a remedial program to be enforced by the state when a school system, taken as a whole, fails to provide a program that meets these criteria of suitability.

So what does this definition of suitable education have to do with diplomas and high school graduation? I believe that requiring accountability for the outlined elements of a suitable program is the best way to certify high school completion. There are four major exit requirements that I believe must be met for each student; the responsibility for these falls predominantly on the school:

1. The school must certify that each student has had equal access to a quality education throughout his or her twelve years of schooling. Every child must be guaranteed protection from discrimination that prevents him from receiving the education he requires.

2. The school system must have provided each student with a broad range of learning opportunities — in both basic and applied skills — that will enable him or her to function successfully now and in future life.

3. The system must have helped its students along the way to reach their full potential. No school system can force a child to learn; but every school system is responsible for aiding and encouraging the child. That means using our vast wealth of knowledge about learning to identify children's talents, abilities, and interests, to uncover learning problems, and to solve them. It is here that tests of many varieties may play an important role.

4. Finally, there must be clear expectations of what students

should accomplish. The schools are responsible for establishing such expectations; the students are responsible for meeting them. The high school diploma should continue to be a personal statement of accomplishment representing participation in certain activities and sufficient achievement in basic academic skills. Ideally, those expectations should be established on an individual student basis. At the very least, they should be decided by the local school system and the local community. They should not be set at the state level. Students, parents, and educators should all agree on the value and significance of the high school diploma and the qualifications for earning it.

conclusion

Many would like to see high school education defined by a single standard, a neat listing of the accomplishments that all graduating students will have, efficiently identified by a single test score. However, I cannot agree. I like to think, as George Bernard Shaw did, that education is "the child in pursuit of knowledge, and not knowledge in pursuit of the child" (Peter, 1978, p. 173). I cannot and will not believe that the ultimate goal of education is achieving a minimum score on a single test. We want everyone to go beyond minimum level. We want each child to reach his or her maximum potential. And each child is different; there is no test that measures that difference effectively.

But there is no need to "cop out" on the accountability issue. I propose, instead, a dual accountability. First, we must hold the schools firmly accountable for opportunities for learning. They must guarantee each student the instruction, evaluation, and special assistance he or she requires. Only in this way can we effectively certify the success of schools. And I believe that if we take care of the first part then the second part – the certification of students – will take care of itself. This does not, however, relieve the students of responsibility; they should be held accountable for their own performances. The students should be fully aware of what is expected of them, and the diploma should be their reward for meeting those expectations. Legally and morally, we owe students the opportunity to learn. But we also owe them what they owe themselves: an expectation of the excellence of which each of them is capable.

reference

Peter L. J. *Peter's Quotations: Ideas for Our Time*. New York: William Morrow, 1977.

Mark R. Shedd is Commissioner of Education, Connecticut State Department of Education.

Programs for determining placement in remedial instruction, certifying competence in reading and writing skills, and evaluating instructional programs illustrate how testing can influence the awarding of college degrees.

testing and the college degree

r. robert rentz

It is not my intention in this chapter to answer the question of whether or not testing should be involved in awarding college degrees. That question is quite compelling, but, unfortunately, I do not know the answer. A much less compelling question, but one that is much more manageable, is: *How* can testing influence who is awarded the college degree? Testing has, of course, always been used by those concerned with awarding college degrees. Probably more tests are administered in college classrooms during the first months of the fall term than are administered over several years by the College Board and the American College Testing program combined, and improvements could certainly be made in many of those classroom tests. However, my concern is not with the testing program that originates with individual professors or even with the faculties in specific departments; rather, I am concerned with the kind of testing that receives its major impetus from outside the faculty—suggested, mandated, or legislated by administrators, governing boards, or state legislatures—and that is used to assess minimum competency for granting diplomas. Numerous examples of these externally mandated testing programs may be cited. At the state level, for instance, are the Georgia program that I will describe shortly, and the program required by the recent Florida law

that calls for the administration of entrance and exit exams for teacher education candidates. Other programs are unique to individual higher education institutions. There is generally substantial faculty involvement in their development and implementation of such programs, but the impetus for them comes from outside the faculty.

I think the primary motivation behind the development of these testing programs is the popular belief that college graduates are simply not as well educated as they should be. While there is little direct evidence of the performance levels of today's college graduates, there is much popular comment. Anecdotes abound about the graduate who cannot write letters of application, memoranda, or even simple sentences. There are other stories about sixth grade teachers who cannot read at the level of their own students. It is difficult for many citizens to believe that four years of college experience will transform a freshman class, over a quarter of whose members must take remedial English and math, into a group of graduates that can function at a level expected of college graduates. Factors such as the necessity for minimum competency testing of high school graduates and generally declining Scholastic Aptitude Test (SAT) scores tend to erode the base of confidence in the ability of the entering college student. Thus, in the absence of evidence to the contrary, the notion that the college graduate is somehow educationally deficient persists.

In 1972, partly as a response to this general uneasiness and partly to gather information for program improvement, the university system of Georgia began a testing program designed to assess the reading and writing skills of college students during their sophomore year. The Georgia system, composed of thirty-three state-supported junior colleges, senior colleges, and universities; quickly discovered that some 25 to 30 percent of its students could not achieve the minimal levels of performance expected of them in the two tested areas. These findings were partly responsible for the establishment of a formal statewide remedial program in all institutions, accompanied by extensive placement testing of incoming freshmen. At the same time, to help individual departments evaluate their programs, the university system inaugurated major area examinations to be given to bachelor's degree candidates at their exit point. Hills (1977, p. 9) calls these entrance and exit testing activities "a very extensive and elaborately coordinated program of testing. No other state has anything quite like it."

The Georgia programs offer illustrations of several functions that testing can perform in awarding the bachelor's degree. In the remainder of this chapter, I will focus on three of these functions: placement, certification, and program evaluation.

testing for student placement

Testing has been used for selection in college admissions for many years. Test scores, along with high school grades and other information, have assisted college admissions officials in making decisions about who would be admitted. When the number of applicants far exceeded the available space, high selectivity was the general practice followed by most institutions. In recent years, however, the opposite has been true; generally, there has been more space than applicants, and the pressure to maintain enrollment levels has necessitated less reliance on previously used selection criteria. As the ratio of space to applicants has changed in favor of the applicants, the type of admissions decision that must be made has also changed: selection decisions have become placement decisions. Selection means deciding whether or not to admit particular students, whereas placement involves determining which level of instruction or type of program is best suited to individual applicants. A rather comprehensive explication of placement options has been provided by Willingham (1974), who emphasizes accommodating individual student differences by matching students with appropriate educational programs. Placement decisions typically involve such options as exemption from particular courses, advanced placement, or the use of remedial programs. This last option becomes increasingly prominent as more of the less-qualified applicants are admitted.

Placement in remedial programs in Georgia colleges involves a two-stage decision that uses test information. All applicants for each of the thirty-three institutions in the university system are required to submit SAT scores. Students with a combined verbal and math score of less than 650 (on a scale of 400 to 1600) are required to be further tested with a set of tests called the Basic Skills Examination. Students who score below an institution's cut-off point on any of the three parts of the Basic Skills Examination—math, English, or reading—must enter that institution's formal remedial program in those areas in which they are deficient. Before exiting, the students in the remedial program must again take and pass the part or parts of the Basic Skills Examination that they previously failed. Students are allowed up to one year to complete these requirements. Those who begin but never complete the remedial programs and still enter the regular college program will never receive degrees. Thus, in this sense, passing the Basic Skills Examination becomes a requirement in itself for obtaining a degree.

While this use of the Basic Skills Examination involves the func-

tion I have chosen to call certification, it suggests certain characteristics of tests of this sort that should be mentioned here. Using the Basic Skills Examination for both placement and certification at thirty-three institutions throughout the state puts a severe strain on both test security and, to some extent, the credibility of the certification process. Such a problem can only be solved by issuing new test forms at fairly frequent intervals. In fact, the requirement for multiple, equated test forms issued on a regular basis is necessary for the successful implementation of most certification examinations, including those in the minimum competency testing movement. In Georgia, we have dealt with the multiple forms problem by abandoning traditional test development procedures, as well as the purchase of off-the-shelf tests, in favor of an item bank approach and the use of latent trait methodology. The Basic Skills Exams are developed locally on the basis of the Rasch model. Rasch model procedures provide simple and efficient solutions to the problems of equating test forms, and they offer the benefits of an item sampling approach to the item analysis task (see Rentz, 1978).

testing for certification

The clearest example of Georgia's use of tests for certification is the Regents' Testing Program, which assesses the reading and writing skills of students during their sophomore year. Passing this test is required for graduation with either an associate or bachelor's degree. The policy of the board of regents of the university system, adopted in 1972, contains the following statements (Board of Regents, 1972, pp. 554-555):

> It is the responsibility of each institution of the University System of Georgia to assure the other institutions, and the system as a whole, that students obtaining a degree from that institution possess the basic competence of academic literacy, that is, certain minimum skills of reading and writing. . . . Students enrolled in degree programs will be required to take and pass the test. . . . Passing the test is a requirement for graduation.

The battery of tests used in the Regents' Testing Program is called the Language Skills Examination. These tests are given four times a year to about 30,000 students. Students are permitted to take the tests as many times as desired, subject to any required remediation policy of the local institution. (Board policy requires the local institution to provide a remedial program for those failing the test, and it

permits the institution to require the student's participation.) The tests in the Language Skills Examination are locally developed. The reading test is a conventional multiple-choice comprehension and vocabulary test, but the writing test is an actual written essay.

The content of these tests is determined by representatives of the faculties; their aim is to define a minimum level of performance that can reasonably be expected of a graduate regardless of the institution attended. In such a large and diverse student population, what proficiencies can be certified and how? Insofar as content is concerned, there are four options: (1) certification on a course-by-course basis, the process currently in common use, in which each professor assesses students' competence by assigning grades; (2) certification of students' competence in their major areas of study, an option in widespread use but usually operated by recognized groups outside the college (examples include state teacher certification boards and other professional licensing and certification boards); (3) certification based on a core curriculum—a common body of content that each student is expected to master and that is very difficult to define; and (4) certification of basic skills, the solution illustrated by Georgia's Regents' Testing Program.

testing for program evaluation

The Regents' Testing Program also serves a program evaluation function. The percentage of students who have passed the Language Skills Examination in each institution in the Georgia system is reported regularly. The results vary widely among schools, which, over the years, has resulted in extensive studies of lower-division programs, particularly English composition. Program evaluation is not its major thrust, although the Regents' Testing Program can lead to changes in programs; the impact its results have had on curricular programs has created mixed feelings about its overall effectiveness.

The Major Area Examinations, sometimes called senior exit exams, represent a testing activity in the Georgia system that can be readily identified with evaluation. These exams are selected, administered, and reviewed by the local institution. Each department selects or devises its own exam, but the tests used most frequently are the advanced tests of the Graduate Record Examinations. Each graduating senior must take a Major Area Examination; psychology majors take a psychology exam, biology majors a biology exam, and so on. However, there are no passing requirements; the results are used by each academic department as part of a review of its academic program. Since this particular testing program is relatively new, its usefulness is yet to be determined.

Considerable testing takes place in the Georgia system, and one thing is clear: testing influences curriculum. It would require considerably more space to describe all the various ways in which these testing programs have influenced the higher education system in the state, but I will mention a few related to the Regents' Testing Program. Course content has changed. New courses have been added. More essay tests are given and more in-class writing is done in other courses besides English. And faculty are increasingly conscious of and concerned about their responsibilities to students in teaching basic skills.

The reactions to these trends have been both positive and negative. One junior college dean writes, "I believe the Regents' Test has done more than any other single device to improve the quality of higher education in this state" (Austin, 1978). Yet the head of an English department (Corse, 1978) declared:

> However, because we now are devoting our best efforts to getting the largest number of students past the essay exam as possible, we are teaching to the exam, with an entire course, English 111, given over to developing one type of essay writing, the writing of a five-paragraph argumentative essay written under a time limit on a topic about which the author may or may not have knowledge, ideas, or personal opinions. Teaching this one useful writing skill has the beneficial effect of bringing large numbers of weaker students to a minimal level of literacy; but, at the same time, it devastates the content of the composition program that should be offering the better student challenges to produce writing of high quality. Because the Regents' Test is primarily designed to establish a minimal level of literacy, our teaching to this test, which its importance forces us to do, tends to make the minimum acceptable competency the goal of our instruction, a circumstance that guarantees mediocrity.

conclusion

In this chapter, I have approached the issue of how testing can be used as a determinant in awarding college degrees by describing several testing programs in the university system of Georgia. These programs illustrate three functions testing can perform—placement, certification, and evaluation. In some ways, these functions influence the individual directly; in others, students are influenced by program changes brought about by the testing. As we have seen, testing can be a powerful agent for change. If we are now facing an era of more widespread use of tests for determining eligibility for college degrees, then

colleges must be aware of the impact such testing is likely to have on their campuses.

references

Austin, M. Personal communication, September 25, 1978.

Board of Regents of the University System of Georgia. *Minutes from the April 1972 Meeting of the Board of Regents*. Atlanta, Ga., 1972.

Corse, L. B. Personal communication, September 26, 1978.

Hills, J. R. *Proficiency Testing: Implications for Higher Education*. Atlanta, Ga.: Southern Regional Education Board, 1977.

Rentz, R. R. "Monitoring the Quality of an Item-Pool Calibrated by the Rasch Model." Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Ontario, March 1978.

Willingham, W. W. *College Placement and Exemption*. New York: College Entrance Examination Board, 1974.

R. Robert Rentz is associate professor of educational foundations and director of the Regents' Testing Program at Georgia State University.

Decisions on what kinds of interventions to to evaluate and what kinds of data to collect are needed for wise use of evaluation resources.

critical decisions in evaluation studies

peter h. rossi

The human services field presently seems to have climbed onto a plateau, and it appears able to gain additional altitude in small increments only by expending considerable additional amounts of effort and resources. Public and health care efforts have lowered the death rate close to its asymptotic minimum and decreased its variance due to social class, ethnicity, race, and place of residence. Further reductions in mortality are going to be difficult to achieve, will require very heavy expenditures, and may increase undesirable side effects. Similarly, we have gone about as far as we can go with our criminal justice system in working to keep crime under control. Further progress may take more effort than we can afford. In education, compulsory school attendance until ages sixteen to eighteen plus the availability of state-supported colleges and universities have brought our country a long way on the road to universal literacy. But ironing out the variations in educational attainment and intellectual functioning that currently exists is clearly a difficult task. In short, the easy problems in all these fields have been met fairly well; the difficult ones still lie ahead. Indeed, the more problems we solve, the more difficult are the dilemmas that still remain.

Corollary to this generalization is the fact that in field after field new interventions we might devise are not going to have spectacular

effects; we are most likely to consider changes that amount to little more than tinkering with existing systems. While their inventors and advocates may tout such changes as fundamental alterations in our human services, they most likely will turn out to be small variations on existing themes. For example, the mass-produced textbook was a fundamentally important innovation, one that lies at the base of our educational system. And innovative materials for programmed instruction can be viewed as simply new types of textbooks—perhaps better than many but not fundamentally different from most and certainly not as different from regular textbooks as those textbooks were from whatever preceded them.

Thus, the changes we try out on our human services are minor, while the problems we must deal with become increasingly intractable. As a result, any innovation is likely to produce effects that are weak and inconclusive at best. It is this likely outcome that underlies the trend toward increasingly rigorous evaluations of interventions. It is no longer plausible to evaluate educational changes through direct inspection, through the judgments of experts, or through the reports of persons experiencing the changes. We have learned that detecting the effects of interventions requires considerable precision in measurement and powerful research designs. Thus, as the problems become more difficult and the interventions become weaker in their effects, the means for detecting them must become finer; acquiring definitive, valid information about interventions requires considerable effort, resources, and expertise—often at levels that appear inappropriately expensive in relation to the role such information might play in policy decisions about the interventions in question.

when not to evaluate

Given an intervention—some new procedure, device, organizational rearrangement, or whatever—that appears promising, what kinds of measurement information might a decision maker need in order to determine whether or not that intervention is worth installing in an educational institution or system? Clearly the amount of information that would be desirable is dependent on two characteristics of the intervention in question. The first of these is the cost involved: expensive interventions, taking into account not only capital and operating costs but nonmonetary costs as well, would call for more and better information than would relatively inexpensive interventions. For example, it makes absolutely no sense to attempt to measure the impact of using plastic rather than steel paper clips, a judgment that should appear obvious to all. The second characteristic is an interven-

tion's underlying potential: an intervention that may cause some harm should be evaluated more carefully than should one that appears to be completely benign. Of course, harm should be viewed quite broadly; it should include possible damage to organizations as well as damage to pupils and other persons in the educational systems involved.

When stated obversely, these two principles have implications that are not ordinarily taken into account. The obverse says that if an intervention is inexpensive and obviously benign, then it is not worth evaluating as to whether or not it has any impact on some particular educational outcome. The cost of obtaining such information may often more than offset its worth. Furthermore, this is just the sort of intervention that is likely to have little impact, and precise estimates of such impact are extremely costly to obtain. For example, providing enough money to persons released from state prisons to enable them to survive for a month or so would not be an expensive intervention, and it would clearly be helpful to prisoners who are usually released with sums between $25 and $50. In addition, such provision might help reduce recidivism by easing the transition to civilian employment. Detecting such effects is likely to be very expensive, although providing the additional money is relatively cheap. As another example, an educational intervention that would make available to high school math students inexpensive hand calculators is probably not worth evaluating with any great precision. Similarly, a federal program that would provide $5 annually to school systems for each child from a poverty-level household is not worth evaluating as far as impact on the students is concerned. The additional funds could not possibly hurt either the school systems or the pupils, and the cost of properly evaluating whether or not such funds had a positive impact on pupil learning would be extremely expensive. Implementing such a program might be a waste of money, but evaluating it surely would be even more of a waste. In short, for inexpensive, clearly benign interventions, some basic errors are acceptable.

An additional kind of intervention also should not be evaluated. Indeed, a very good case can be made for the belief that such interventions should not even be attempted. I refer to "black box" interventions—the kind for which no specified rationale, theory, or model postulates how the intervention is going to accomplish its aims. Perhaps the most frequently employed black box interventions are those that involve giving unearmarked funds to school systems or schools in the vague hope that they will somehow improve themselves. But many other interventions are proposed as well, such as Head Start

programs and educational vouchers, whose specific mechanisms for producing effects are left unstated.

Black box interventions often arise out of what may be called failure of nerve. Often, despairing of finding any specific intervention for which some sound rationale can be proposed, an intervention is devised that primarily consists of providing incentives for innovation. Thus, a school system might turn over direction of local schools to neighborhood school boards in the optimistic, democratic belief that local parents may be better able to specify a school curriculum than are system-wide school boards. There may be good reasons for decentralizing control of schools, but the idea that such moves will somehow improve the schools because parents are better at making educational decisions than are educators is not one of them. A similar example of failure of nerve is subcontracting to private firms to provide instruction under a profit-incentive system. Here the rationale is simple, if unsatisfactory: in the absence of any notion of what to do to improve the schools, simply provide profit incentives for improvement, and improvements will appear. The contract learning experiments sponsored by the Office of Economic Opportunity (Gramlich and Koshel, 1975) demonstrated how futile such attempts were in producing startling or effective innovations, at least in the short run.

The main reason that black box, failure-of-nerve interventions are not worth evaluating is that we learn so little from doing so. Evaluation of a well-thought-through intervention with specific goals and clear means for reaching them provides decision makers with valuable information about what to do next if the intervention fails or is only marginally successful. In the case of a black box intervention, however, since the mechanism of success is unknown, we are unable to sort out the essential from the inessential aspects of the intervention and hence will likely be unable to reproduce or enhance desirable effects in other settings. Evaluations of such interventions characteristically provide go-or-no-go information and do not add cumulative knowledge bases. There simply is no substitute for general understanding and theory in the design of interventions. Mindless innovation may produce some movement, but it will result in little progress.

questions an evaluator should answer

The considerations raised so far, in sum, add up to a positive definition of an evaluable intervention—that is, one that is worth the funds that must be expended in order to determine with some precision whether or not it is having its desired effects. Evaluable interventions can be defined as those that have clearly defined (and measur-

able) goals, are theory-based, would require a heavy investment of fund to operate as a program, and may potentially inflict some harmful effects on persons or organizations. The more of these qualifications a proposed intervention has, the more carefully it should be evaluated. And, conversely, the fewer such qualities are present, the less worthwhile it is as an intervention and the less worthwhile evaluating its effectiveness may be.

If a proposed intervention meets these qualifications, how can its subsequent evaluation best provide useful information to decision makers? The solution requires answers to three interrelated evaluation questions: (1) Is the intervention effective to a significant degree in achieving its goals without substantial negative effects? (2) Can the intervention be delivered (or implemented) successfully within the organizational context in which it would be embedded? (3) Can the intervention produce benefits that justify the costs—both monetary and nonmonetary—that are necessary to achieve its intended effects?

Is an Intervention Effective? Since evaluation research came into fashion ten years ago, we have accrued sufficient experience with applying powerful research designs under field conditions to learn how to discern the effectiveness of interventions; typically, the main obstacles to applying such research to the evaluation of interventions are time and money. We have learned that it is both possible and feasible to carry out quite elaborate randomized controlled experiments under field conditions. We have also learned through experience that, with proper safeguards, nonexperimental statistical methods can also be used with considerable confidence in assessing the impacts of interventions.

In short, it is possible, provided that we are willing to spend the time and can afford the costs, to obtain quite precise, unbiased estimates of the effects of interventions. But there is a difference between statistical proof and its implications for policy decisions. The best the researcher can do is provide data showing that an intervention does or does not produce statistically significant effects. Often, however, when seen in the light of policy needs, statistically significant effects are not important. For example, the Educational Testing Service's (ETS) evaluation of *Sesame Street* (Ball and Bogatz, 1970) showed that children who had viewed that program had progressed farther toward an understanding of certain basic relationships, could recognize more letters of the alphabet, and had a clearer understanding of some rudimentary arithmetic operations than had children who had not seen the program; however, there still remained the question of whether or not such results were significant from a policy viewpoint. The ETS evaluation showed that after the end of a year's viewing, on the average,

viewers could recognize two more letters of the alphabet than could those who did not look at the program. But do two additional letters represent an increase in learning significant enough to justify the effort that went into the design of the program? A somewhat similar case involves the Head Start program. Whether or not it was effective is apparently a controversial issue among evaluators; however, even if we accept the most optimistic of the several findings, can we say whether or not its positive effects are significant enough to merit policy attention?

Furthermore, deciding policy significance involves making judgments as to whether or not unintended side effects cancel out positive primary effects. In his reanalysis of the *Sesame Street* results, for example, Cook (1975) found that the program had stronger effects among middle-class children than among poor children; the end result was a general widening in the learning gap between the two socioeconomic levels of viewers of the program. Similarly, in Seattle and Denver income maintenance experiments, a slight work disincentive effect was shown to result from income maintenance payments, especially among secondary workers in households: mothers of young children, and adolescents. This effect was statistically significant, but its policy implications were not clear; though at first glance appearing to be a major drawback, in some respects—mothers of young children withdrawing from low-paying jobs to keep house and rear their children and adolescents remaining in high school until graduation—it could be judged as a positive outcome. In addition, it was found that the payments fostered the breakup of marriages; this, too, may at first seem negative, but the payments may have provided sufficient income security to free women from unhappy marriages (Hannan, Tuma, and Groeneveld, 1978).

In short, while the researcher can now feel confident that his measurements can provide precise and unbiased estimates of the effects of interventions, this information may not be relevant as far as policy is concerned. Policy significance is not equivalent to statistical significance. Judgments still must be made about the appropriateness of the magnitude of the effects and whether or not there is a satisfactory tradeoff between positive results and negative side effects.

Can the Intervention Be Delivered Successfully? The most useful estimates of intervention effects result from randomized controlled experiments; these must be run by researchers who carefully implement interventions under conditions that ensure its delivery to appropriate target groups. Effects, then, are best measured when an intervention is delivered in a standard way at its intended full strength. For this reason, randomized experiments on a grand scale have been primarily concerned with transfer payments as interventions (these might

include income maintenance payments, subsidized health insurance, housing allowances, unemployment benefits for released prisoners, or similar aid). The delivery of transfer payments is fairly readily evaluated because it can be conducted within the framework of a randomized experiment that simulates or imitates closely the ways in which such payments would be delivered when embodied in a statutory program. Human services delivery, however, is not so easily measured; unfortunately, many human services interventions that work quite well in randomized controlled experiments administered by researchers often fail miserably in other contexts. This is primarily because it is difficult to standardize the delivery of human services, especially when the deliverers are professionals who have considerable autonomy in the exercise of their professional functions, largely as a consequence of the failure of interventions that work well under highly controlled situations to work at all in the field or institutional context.

The expectable difference between pilot runs and production runs often means that an intervention must be tested twice. It is tested first within the context of a carefully controlled experiment; results from such an experiment provide estimates of an intervention's effectiveness under the most favorable circumstances as administered by the dedicated designer of the intervention. It is then tested within the context of the institution that will be given the delivery mandate if the intervention is incorporated into statutory policy.

Perhaps most representative of the sort of double testing suggested here are the experiments sponsored by the Department of Labor (Rossi, Berk, and Lenihan, forthcoming) concerning the efficacy of extending unemployment insurance benefits as postrelease financial aid to ex-felons. Such aid, of course, is intended to reduce recidivism by easing the transition to civilian employment. This intervention was tried initially on a small scale in Baltimore as a randomized experiment; it was run by a devoted social researcher who administered payments and provided job placement services with the aid of a small but conscientious staff. The Baltimore experiment produced very encouraging results, reducing arrests on property-related charges (that is, burglary, robbery, and larceny) by about 8 percent, a hefty 25 percent reduction in recidivism for such charges during the postrelease year as compared with the control group. The Department of Labor then tested the same program in more policy-relevant settings by having the departments of corrections and unemployment security in Georgia and Texas administer it on a trial basis as a randomized controlled experiment. As administered by those agencies, the intervention had no significant impact on arrests on property-related charges during the postrelease year in either state. Under some condi-

tions, financial aid was apparently effective—but not when it was delivered by the kinds of agencies that would be responsible for it if it were enacted as a national policy.

Program implementation involves other measurement issues besides evaluating the effectiveness of interventions. The success of a program depends on how effectively it is implemented. Hence, measuring service delivery is a problem in the administration of all programs, especially those, such as human services, that must rely heavily on personnel for delivery. From the perspective of decision making, it is necessary to know not only whether or not an intervention will work under certain circumstances but also whether or not it will work within the context of the institution that will have to administer it. This suggests that policy making should be more tentative in establishing a program, making provisions for both close tracking of how well the program is being implemented and periodic checks on its effectiveness as it is delivered.

Do the Intervention's Benefits Justify its Costs? The general idea behind benefit-to-cost analyses is quite simple: policies that create benefits greater than their costs are the only ones worth enacting, and policies with high benefit-to-cost ratios make better use of resources than do those with lower ratios. Going beyond this general idea to the calculation of benefit-to-cost ratios, however, one leaves a simple world and enters a maze of intricate complications. To begin with, benefits and costs may be regarded from many viewpoints—from those of individual recipients of an intervention to those of individual taxpayers, of the institution involved, and finally, of the government administration or the society as a whole. Very costly educational interventions may offer very high benefit-to-cost ratios to recipients but fractional ones to every other party.

Second, calculating a benefit-to-cost ratio requires reducing all benefits and costs to some common metric—usually monetary—units. This may make sense in the calculation of benefit-to-cost ratios for dams and irrigation systems, whose main effects may be calculated in monetary terms, but how can we measure how much a person benefits from learning more math? What is the benefit to society of raising the national average of math scores on the Scholastic Aptitude Test (SAT) by two or three points? It is clear that there are some societal benefits, but it is difficult, if not impossible, to measure such benefits in terms that would make sense to everyone concerned.

Finally, benefit-to-cost ratios are generally very sensitive to the discount rates applied to expenditures. Since investing monies at a given time on an intervention means that alternative investments cannot then be made that might accrue interest over the future, it is

necessary to compare the worth of the present expenditures discounted for the future worth of investment alternatives. Discount rates are largely conjectural, and, for interventions that call for a fairly large amount of present-day expenditures, benefit-to-cost ratios can vary widely.

For these reasons, benefit-to-cost calculations, at least as applied to social programs, tend to approximate the truth value of science fiction—they are interesting, perhaps even insightful, but they are mainly the product of some fertile imagination. This is particularly true of benefit-to-cost calculations applied to programs whose effectiveness has not yet been tested but is simply taken for granted. Costs and benefits, of course, should not be ignored. Indeed, I hold the contrary view. Calculations of cost-effectiveness—that is, the cost of a delivered unit of effectiveness—are especially useful. For example, given a program that is effective in raising the average scores on some standardized test of reading ability, it is possible to compute how much each unit gain in reading scores costs. As a further illustration, Cook (1975) reports that, by his calculations, each additional letter of the alphabet learned by a preschool child through exposure to *Sesame Street* costs approximately $.25. And, in the Baltimore experiment conducted by the Department of Labor, it costs about $12,000 to avert each incident of recidivism, an amount that may seem excessive until one compares it with the costs of processing an arrested person through the criminal justice system and maintaining that person in jail for a typical two-year sentence.

Calculating cost effectiveness requires close monitoring of costs and units of services delivered, as well as measures of effectiveness. The same research operations and measures—with the addition of cost accounting—that can be used to monitor the delivery of services can provide the basic information used to calculate cost effectiveness.

conclusion

This chapter has examined some of the major issues that arise in measuring the effectiveness of and making decisions about interventions. Assuming that precise and accurate measurements of the effectiveness of interventions are expensive, I have stressed that there are circumstances under which one should not undertake measurement: some interventions are simply too trivial to waste resources on, and others are so poorly defined that any measurement is bound to be baffling and equivocal. For those interventions that are evaluable, I have illustrated the considerable difference between the effectiveness of an intervention conducted under pilot-run conditions and the effective-

ness of one administered by the institution that will have the ultimate responsibility for it if the program is enacted. I have also expressed a pessimistic view of benefit-to-cost calculations as being largely conjectural and ordinarily highly dependent on shaky assumptions. In their places, I have stressed the usefulness of measuring cost effectiveness.

Measuring the effectiveness of interventions and the costs associated with measured effects provides only one part of the information that goes into the decision-making process. No matter how well an evaluation is conducted, it would be naive to expect the resulting measures of effectiveness to have an all-determining impact on the decisions of policy makers. There are many reasons to enact interventions into policy without considering their effectiveness. For instance, equity considerations may completely outweigh considerations of effectiveness. In addition, constituency demands arising from clients, organizations, and perhaps even suppliers may appear more cogent to decision makers than the representations of evaluation researchers. Indeed, would one have it otherwise? In a democratic society, is it not better to have policy that is responsive to the push and pull of politics than to the outcomes of social research?

references

Ball, S., and Bogatz, G. A. *The First Year of Sesame Street: An Evaluation*. Princeton, N.J.: Educational Testing Service, 1970.

Cook, T. D., and others. *Sesame Street Revisited*. New York: Russell Sage Foundation, 1975.

Gramlich, E. M., and Koshel, P. P. *Educational Performance Contracting: An Evaluation of an Experiment*. Washington: Brookings Institution, 1975.

Hannan, M. T., Tuma, N. B., and Groeneveld, L. P. "Income and Independence Effects on Marital Dissolution: Results from the Seattle and Denver Income-Maintenance Experiments." *American Journal of Sociology*, 1978, *84* (3), 611–633.

Rossi, P. H., Berk, R. A., and Lenihan, K. A. *Money, Work, and Time*. New York: Academic Press, forthcoming.

Peter H. Rossi is director of the Social and Demographic Research Institute at the University of Massachusetts.

Responses of policymakers to Title I and bilingual education program evaluations illustrate how research affects policy decisions.

using measurement in educational decision making

john ellis

Throughout the last decade, and most certainly during the past few years, lawmakers and educational administrators in the federal government have relied increasingly on measurement and evaluation in making their decisions. The reasons are clear enough: a growing demand for accountability in education and a clear need for more accurate yardsticks to measure the efficacy of educational systems, programs, curriculums, and student learning. Congress wants more objective, pragmatic evidence on which to base its decisions about whether support for programs should be increased, decreased, or abandoned. Educational administrators, teachers, parents, Office of Management and Budget staff, and others outside the legislative halls also want objective evidence to support their proposals and programs.

My current role involves working with all 120 programs administered by the U.S. Office of Education (OE). In this capacity I can see, though somewhat dimly, the constellation of forces that focus on the congress and the administration and attempt to persuade, cajole, lure, or threaten them into taking appropriate action. In this chapter, I will examine the influence of testing and research on two of OE's major

concerns—Title I of the Elementary and Secondary Education Act (ESEA) and bilingual education. These two areas exemplify the increasing impact testing is having on decision making at the federal level; in addition, they show how testing may be applied in other educational program areas.

testing and Title I

Title I of ESEA is the flagship of federal elementary and secondary education programs. It was authorized in 1965 to provide special educational services to educationally deprived children in low-income areas, and its budget has grown from $959 million in fiscal year 1966 to over $3 billion today. Title I funds now go to 14,000 of the 16,000 school districts in the country. Measurement, in the form of evaluation, has been an integral part of Title I since its enactment. However, first-generation evaluations were basically efforts to find "successful" Title I programs. By today's standards, they were relatively primitive and imprecise, and they were inadequate and unsatisfactory for the purposes of the OE and Congress. In fact, there were serious discussions about whether to perform radical surgery on Title I or to abandon it altogether.

In 1974, Congress amended Title I by adding several new duties for the U.S. Commissioner of Education. The commissioner was to: strengthen the requirement for independent evaluations of Title I programs and projects; develop and publish standards for evaluating the effectiveness of those programs and projects; consult with states to provide jointly sponsored, objective evaluation studies; provide states with evaluation models, utilizing objective criteria and methodology to produce data that are comparable on a statewide and nationwide basis; and provide states with technical assistance for developing and applying their evaluation programs.

As a result of the 1974 education amendments, the National Institute for Education (NIE) and OE conducted a number of studies and surveys. Contrary to earlier findings, these studies showed an increase in achievement for Title I children. Moreover, the NIE study indicated that the effectiveness of Title I programs directly correlated with the quality of administration: programs that were administered well tended to be better than poorly administered programs. These studies led to the general conclusion that Title I was indeed working and that it could be made even better.

Congressional action in 1978 reflects, in part, the results of the NIE and OE studies, for a significant increase in Title I appropriations

was made. Including the new concentration proposal at $400 million, the total Title I appropriation will be over $3.4 billion; this new total represents an increase of 27 percent over last year, the largest increase in the history of the program. Furthermore, Congress placed more responsibility on the states for monitoring Title I programs and incorporated into law some of the provisions in current regulations dealing with program administration.

The research conducted by NIE was also good for the agency. The report of the Senate/House Committee on the reauthorization of ESEA stated, "The high quality and extremely useful work accomplished by NIE in the ESEA Title I study was particularly influential in impressing the Congress that the institute has grown and matured. It now represents a unique and solid resource [that] administrators and educational policy makers can depend on for the study of difficult and previously unknown areas [that] affect learning and the education process, as well as national education policy issues" (Conference Report on H. R. 15, 1978, p. H12224).

The education amendments of 1978 reflect Congress' desire for still more and better measurement. A lengthy section on program evaluation specifies that the commissioner of education shall continue to provide for independent evaluations of Title I programs and projects as well as technical assistance. In addition, the commissioner must report the results of evaluations to Congress no later than February in 1980, 1982, and 1984.

A number of Title I evaluation studies are currently under way; three that should be completed by next spring undoubtedly will have substantial impact on Title I legislation next year. Those conducting these studies are seeking to determine: (1) what percentage of students retain fall-to-spring achievement gains during the summer; (2) the cost effectiveness of the various types of Title I services; and (3) the nature and extent of parental involvement in the education of children.

Talk about the failures of Title I has virtually disappeared. There is no question that constituency pressures and social needs overshadowed any test results in determining funding levels for Title I. Yet Congress clearly wanted to ensure that the dollars appropriated were being used wisely: congressional committees continued to press OE for evidence that the programs were working. But in this session, Congress turned away from questioning the desirability of having such a program, instead focusing its attention on making Title I more flexible and more effective. Without data documenting student success and pointing the way toward program refinements, I seriously doubt that such positive congressional action would have been taken.

evaluation and bilingual education

An estimated 15 million persons of limited English-speaking ability live in this country. About 24 percent of them, or 3.6 million, are four to eighteen years of age and therfore of particular concern to our public and private schools. An overwhelming number—69 percent, or 2.1 million—of these young people speak Spanish. Only five other languages account for more than 50,000 persons each: Italian, French, Filipino, German, and Chinese.

In 1968, Congress enacted the Bilingual Education Act as Title VII of ESEA and appropriated $7.5 million for bilingual education. In 1974, the U.S. Supreme Court, in *Lau v. Nichols,* ruled that the San Francisco school district must provide special programs for children of limited English-speaking ability. Although the court did not specifically require bilingual education, that approach was one option for meeting the new requirements and assuring equal access to education. Following that decision, Congress substantially broadened Title VII to help states and school systems better serve non-English-speaking students. New amendments called for more deliberate and systematic teacher training and curriculum development. They also authorized funds for creating resource centers to help teachers, as well as materials development centers and assessment and dissemination centers.

Unfortunately, research in bilingual education to date is fragmentary and inconclusive. A major study of the subject was conducted by the American Institute of Research (AIR) under a $1.5 million contract with OE. Results of this study, released in the spring of 1977, caused reverberations in the educational community that are still being felt today. They found that less than one third of the students participating in bilingual classes were of limited English-speaking ability and that, in the judgment of teachers, approximately three fourths of the fourth, fifth, and sixth graders in Title VII classrooms were either English monolingual or English-dominant bilingual students (Danoff, 1978). The researchers also noted that, in their study sample, Title VII students had slightly lower grades in English than did students who were not in Title VII programs; in mathematics, across grades, they were performing at about the same level as students not in Title VII.

In August 1978, the National Conference on the Education of Hispanics issued a statement saying that the AIR report had been "seriously questioned by several independent researchers of renowned competence." The conference went on record as repudiating the report and passed a resolution asking that OE also repudiate the report and

take steps to replicate the study. Others have made similar requests (U.S. Office of Education, 1978).

It is understandable that the Hispanic community would be upset. When a research study that seriously questions the results of a program designed to address some of the long-neglected cultural, linguistic, and academic concerns of Hispanics, it is to be expected that such a study would itself be subject to intense scrutiny. It was not unlike the early days of Title I when negative results appeared so frequently.

Congress, of course, was concerned about claims that the majority of pupils in the program were competent in English. Thus, when it acted on the bilingual program this year, it mandated that no more than 40 percent of the pupils in the program should be children whose native language is English. (To avoid problems of segregation, some English-speaking pupils had to be eligible to participate.) Congress also changed the description of these children from "limited English-speaking" to "limited English proficiency," since speaking is only one factor that should be considered. In addition, Congress increased its appropriation for bilingual programs to $150 million and called for additional research. The research that will emerge, including studies on entry criteria for bilingual education programs, exit criteria, program effectiveness, and teacher training will have an impact on future appropriations. Clearly, the research available is insufficient for making important decisions.

national testing

A final concern I wish to address in this chapter is the alleged specter of a federally sponsored national competency test. The Carter Administration has expressed its opposition to such a federal role. Joseph A. Califano, Jr., Secretary of Health, Education, and Welfare, agreeing with a report from the National Academy of Education, stated: "A national test would improperly centralize a matter of state and local control" (Califano, 1978, p. 4).

What is the federal role? I believe it should be based on how we can best help state and local school districts. In the session just adjourned, Congress authorized the U.S. Commissioner of Education to make grants to states and to individual school districts for implementing educational proficiency standards and providing assistance with achievement testing. While no money for this program has yet been appropriated, the debate about this legislation is instructive. I recall hearing late at night the conference-committee dialogue con-

cerning the issue of federal control. To guard against federal influence, a protective statement was adopted concerning proficiency standard assistance that says: "Nothing in this section shall authorize the commissioner to impose tests on state educational agencies or local educational agencies, and no such agency shall be compelled in any way to apply for funds under this section" (Conference Report on H. R. 15, 1978, p. H12181). Similarly, regarding assistance with achievement testing, the safeguard provision stated: "Nothing in this section shall authorize the commissioner to require specific tests or test questions. Any state or local educational agency may refuse to use any test or test question developed under this section" (Conference Report on H. R. 15, 1978, p. H12181).

Congress is responding to public pressure and test results concerning achievement levels in American schools. However, sensitive to the dangers of federal control, it has placed clear limitations on the Office of Education in administering the laws.

In summarizing this brief trip through some recent decisions, I would make the following general conclusions:

1. In an increasingly complex society, tests will continue to have an important impact on individuals, institutions, and the decisions that are made about education.
2. Constant vigilance must be exercised to ensure that the federal role continues to be one concerned with research, technical assistance, and funding rather than one of domination or control.
3. The educational research establishment must expand its methodological approaches from traditional reliance on psychology and statistical analysis to include the use of the wider range of methodologies now common in other sciences.
4. Policy makers at all levels must be willing to make intelligent adjustments to programs based on results.
5. While increased dollars and continuing authorizations are welcome signs, we must remind ourselves that the real measures of success are how well students learn and how significantly their life chances are improved.

conclusion

The short political life cycle of people and events in Washington often stresses instant success, but it should become increasingly apparent that, in the long run, the best policy will be to support programs that demonstrate positive and tangible long-term results. It is naive to believe that research, however sophisticated, will resolve

intensely political questions. Conversely, it is unnecessarily cynical to believe that research results have little or no effect on legislative action. Congress and the administration do take seriously responsible evidence of program effectiveness, and I sense a willingness to make effective use of the results of major studies. Critical questions remain unanswered, and numerous decisions must be made about the focus of programs and the allocation of scarce resources. Solid research aided by refined measuring instruments and new methodologies will be increasingly helpful in making those decisions in the years ahead.

references

Califano, J. A., Jr. "Remarks of Secretary Joseph A. Califano, Jr. to the Conference on Achievement Testing and Basic Skills." Speech delivered to the Conference on Achievement Testing and Basic Skills, Washington, D.C. March 1978.

"Conference Report on H. R. 15." *U.S. Congressional Record*, October 10, 1978, pp. H12136 12224.

Danoff, M. N., and others. *Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program-Overview of Study and Findings.* Palo Alto, Calif.: American Institute for Research, 1978.

U.S. Office of Education. *Recommendations and Resolutions of the National Conference on the Education of Hispanics.* Washington, D.C.: U.S. Office of Education, 1978.

John Ellis is Executive Deputy Commissioner for Education Programs, U.S. Office of Education.

index

97

F

G

H

I

J

K

L

M

from the introduction

Demands by educational policy makers for applications of measurement to significant new tasks are having far reaching effects on education and measurement. Measurement professionals are being asked to help in designing educational programs for, among others, children with learning disabilities, gifted children, and bilingual children. These professionals are also being asked how measurement can help in allocating funds to schools, determining qualifications for high school diplomas and college degrees, and evaluating the worth of new educational programs. These new and complex demands have given rise to congressional debate, federal and state conferences, and extensive discussion and developmental work by measurement specialists. Measurement and educational policy is the theme of this inaugural volume of New Directions for Testing and Measurement, *which includes the ten papers presented at the 1978 Educational Testing Service Invitational Conference.*